FINDING OURSELVES:

THOUGHT-EXPERIMENTS AND PERSONAL IDENTITY

SIMON BECK

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Philosophy

UNIVERSITY OF CAPE TOWN

February, 1994

CONTENTS

ACKNOWLEDGEMENT

The financial assistance of the Centre for Science Development (HSRC, South Africa) towards this research is hereby acknowledged. Opinions expressed in this work, or conclusions arrived at, are those of the author and are not to be attributed to the Centre for Science Development.

ABSTRACT

The central concern of this thesis is with the role thought-experiments play in the debate about personal identity, especially with the question of what role they should play.

The thesis is divided into two parts. The first part is a defence of the use of thought-experiments against a number of influential and potentially damaging indictments of it. Some of the arguments discussed are directed at specific experiments or a specific kind of experiment, but all have implications which extend to the method in general. The thrust of my response to these arguments is that even if some objections to thought-experiments are strong enough to make us more cautious about how we use them, none of them is strong enough to require the general abandonment of the method of thought-experiment in the context of the personal identity debate.

The aim of the second part is to find an answer to the question of what it is that thought-experiments can do, given that there is no prior case ruling them out altogether. The strategy is to reach an answer by a close examination of some prominent examples of thought-experiments in the literature.

In the nature of my topic, there are two issues here. One is methodological, about what one can expect from a thought-experiment; the other is the substantive one as to what thought-experiments can really establish about the nature of personal identity.

With regard to the methodological issue, two basic kinds of potentially informative thought-experiment emerge. There are those which serve to support or undermine a theory by revealing the relative importance of the various principles of classification which are implicit in our use of the concepts of person and personal identity. There are also those which function to show that a theory suffers from internal inconsistencies or that it has unacceptable consequences.

In the process of investigating how thought-experiments can work, I argue that one view of personal identity receives stronger support from them than any of its rivals. This is a nonreductionist view which holds that while personal identity can be analysed in terms of psychological continuity, it cannot be <u>reduced</u> in the standardly accepted sense of that term.

CHAPTER 1: INTRODUCTION - THESE BIZARRE FICTIONS

SECTION 1: *Thought-experiments and personal identity*

> For should the soul of a prince, carrying with it the consciousness of the prince's past life, enter and inform the body of a cobbler, as soon as deserted by his own soul, every one sees he would be the same person with the prince, accountable only for the prince's actions: but who would say it was the same man? (Locke 1694: II,xxvii,15)[1]

The modern debate about personal identity begins with Book II, Chapter xxvii of Locke's Essay Concerning Human Understanding. The use of thought-experiments does not begin with the work of Locke[2]. Nor is it restricted to the discussion of persons and their identity, but the role that thought-experiments like the one above have played in that discussion is not paralleled anywhere else in philosophy. Their presence infuses the entire debate: they are used by advocates of all the central competing theories, and all the watersheds in the history of the debate are marked by their use[3].

These points are enough to establish the importance of this method of argument to the topic of personal identity, and to make an enquiry into its workings a worthwhile philosophical enterprise. But there is much more besides this which makes the enquiry a pressing one. The past forty years have seen the emergence of a number of sharp and far-reaching criticisms of the use of thought-experiments, and the clamour of this criticism has recently been on the increase. Given this growing opposition, together with the influential role in which they have been cast, the whole problem will need careful attention if the personal identity debate in its current form is to retain its central position in contemporary metaphysics.

In this introductory chapter I will give a brief sketch of the debate itself and the place thought-experiments have had in it. I will also outline how I will structure my investigation of the method and the place it should have in the personal identity debate.

SECTION 2: *Early historical examples of the personal identity debate*

What is known as the personal identity debate consists, as do most such prominent philosophical debates, in not one but a whole number of debates. Nevertheless, those participating can be divided without much misrepresentation into two overall groups. There are those who think that the

relation of being the same person over a period of time can be analysed into more familiar and better-understood relations, and those who deny that this is the case. The terms used for each group differ in the literature. The most common distinction is drawn at present between "reductionists" and "nonreductionists" (for example, by Parfit 1984); sometimes the divide is expressed as separating "empiricists" from "non-empiricists" (Madell 1981) and sometimes "materialists" from "dualists" (Swinburne and Shoemaker 1984).

These different titles serve to give some idea of who the combatants are and perhaps something of what is at stake, but none of them fits perfectly. While most dualists do tend to see personal identity as unanalysable, there are materialists who also see things this way. There are also those, like Locke himself, who are neutral in the debate as to whether or not mind is matter (Locke 1694: IV,iii,6), but who take up a strong position as to the analysability of personal identity. While it is common for empiricists to insist that the relation can be further analysed in terms of observable relations, there are empiricists who believe that it cannot, and non-empiricists who believe that some analysis is possible. The reductionist/ nonreductionist distinction is perhaps the most accurate of the three standard characterizations, but even so, there are nonreductionists who accept the analysability of personal identity, as we shall see in Chapter 7. As a result, while it may make the exposition more clumsy, I will tend to stick to talk of analysability unless the context dictates

otherwise, but the relevance of the other distinctions should become clear in due course.

The fundamental opposition outlined above can be well illustrated from the early history of the debate, as can the role of thought-experiments in the support of both sides. As I have said, the first major contribution was that of Locke. Locke suggests in his Essay that the notion of "same person" - the notion crucial to law and morality which underlies the institutions of responsibility, commitment, desert, and so on - can be analysed into a more fundamental relation:

> as far as any intelligent being can repeat the idea of any past action with the same consciousness it had of it at first, and with the same consciousness it has of any present action; so far it is the same personal self. (Locke 1694: II,xvii,10)

His response to those who oppose the analysability of the relation takes the form of an attack on the Cartesian notion of the self being an immaterial substance, the focus of the attack being on substantiality rather than immateriality (Jolley 1984: 130). In this, he makes use of a thought-experiment: he holds that were someone to insist that his soul were that of Socrates - that is, he and Socrates were the same immaterial substance - while being unable to remember any of Socrates's actions or thoughts, we would deny that he and

Socrates were the same person. And we would still do so even if it could somehow be proved that there was an immaterial substance shared between them (1694: II,27,14).

The other side of the opposition is represented by Leibniz, especially in his New Essays on Human Understanding. Leibniz's position has some complexity to it and shares certain points with Locke's (Curley 1982, Jolley 1984), but one central strand concerns the importance of immaterial substances. He challenges Locke with the following thought-experiment:

> Here is something we could much more fittingly suppose: in another region of the universe or at some other time there may be a sphere in no way sensibly different from this sphere of earth on which we live, and inhabited by men each of whom differs sensibly in no way from his counterpart among us. Thus at one time there will be more than a hundred million pairs of similar persons, i.e. pairs of persons with the same appearances and states of consciousness. God could transfer the minds, by themselves or with their bodies, from one sphere to the other without their being aware of it; but whether they are transferred or left where they are, what would your authorities say about their persons or 'selves'? Given that the states of consciousness and the inner and outer appearances of the men on these two spheres cannot yield a distinction between them, are they two persons or are they one and the same?...Since according to your theories consciousness alone distinguishes persons, with no need for us to be concerned about the real identity or diversity of substance...what is to prevent us from saying that these two persons who are at the same time in these two similar but inexpressibly distant spheres, are one and the same person? Yet that would be a manifest absurdity.
> (Leibniz 1765: 245)

Thus, according to Leibniz real identity is not to be expressed in more basic terms, especially not in terms of some observable relation. It is a relation sui generis, simple and unanalysable; and that point is established by a thought-experiment.

SECTION 3: *Contemporary examples of the personal identity debate*

The debate over the possibility of further analysis of the concept of personal identity does not end with the exchange between Leibniz and Locke, nor does their method of conducting it. The most influential contemporary writer on the topic is Parfit, who styles himself a reductionist. Personal identity, he holds, can be analysed as a relation of psychological continuity in a 1:1 relationship; what is more, the world could be completely described without any reference to persons or personal identity (Parfit 1984: 212). This is a more sophisticated account than Locke's, while showing a strong resemblance to it. His support for it also resembles Locke's, with added sophistication. He supports it by attacking those who believe that personal identity is an irreducible relation, focusing on the fact that they also believe that there must always be an answer to any question of

identity. Parfit takes the following thought-experiment to show that there are not always answers.

> My body is fatally injured as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me. And he has a body that is very like mine. (Parfit 1984: 254-5)

In this scenario, all the possible answers to the question as to the identity relation between the original and the survivors are equally implausible. Both cannot be Parfit on pain of absurdity, for then two distinct people would be one and the same. To say that neither is Parfit is tantamount to saying that Parfit is dead, which would likewise be absurd. As he puts it, this would be to call a double success a failure. But the only other option is that one of the survivors and not the other is Parfit, and such an answer would be arbitrary or worse. So nonreductionism must be wrong.

Madell is one modern-day combatant in the Leibniz tradition. His insistence on the unanalysable nature of the relation has behind it these considerations:

> the following two thought-experiments are equally intelligible: (a) that I might not have existed, but someone having exactly the life that I have had might have existed instead; (b) I might have had a totally different life, even to the extent of being born centuries earlier. (Madell 1981: 79)

So the opposing positions and the characteristic method are to be observed at both ends of the modern debate, at the beginning and at the latest point. But the method also marks high points along the way. It was remarked earlier that the debate was not one but many debates, while the exposition so far may create a different impression. For within the camps, especially the camp favouring analysis, there is much dissension, and here again thought-experiments are prominent in the armouries of almost all who are engaged.

The most important difference among those who favour analysis occurs between those who hold that identity is to be analysed as psychological continuity, and those who opt for physical continuity. For Parfit, as we have seen[4], when the variables stand for persons, for x to be identical to y, x and nobody else must be psychologically continuous with y (Parfit 1984: 262-3). For Williams, amongst others, x must be physically continuous with y. Why should we believe Williams rather than Parfit? Because of the following case, Williams argues.

> All the events (Charles) claims to have witnessed and all the actions he claims to have done point unanimously to the life-history of some one person in the past - for instance, Guy Fawkes. Not only do all Charles' memory claims that can be checked fit the pattern of Fawkes' life as known to historians, but others that cannot be checked are plausible, provide explanations of unexplained facts, and so on. (Williams 1956: 7-8)

Should we follow the suggestions of Locke and Parfit, we would agree that Charles <u>was</u> Guy Fawkes, but

> if it is logically possible that Charles should undergo the changes described, then it is logically possible that some other man should simultaneously undergo the same changes; e.g. that both Charles and his brother Robert should be found in this condition. What should we say in that case? They cannot both be Guy Fawkes; if they were, Guy Fawkes would be in two places at once, which is absurd...We might instead say that one of them was identical with Guy Fawkes, and that the other was just like him; but this would be an entirely vacuous manoeuvre, since there would be <u>ex hypothesi</u> no principle determining which description was to apply to which...The only case in which identity and exact similarity could be distinguished...is that of the body - 'same body' and 'exactly similar body' really do mark a difference. Thus I should claim that the omission of the body takes away all content from the idea of personal <u>identity</u>. (Williams 1956:8-10)

Thus on Williams's account the only way to determine who really was Guy Fawkes would be to discover which of Charles's and Robert's bodies were that of Guy Fawkes, and personal identity consists in bodily, not psychological continuity.

There are also arguments to support Parfit against Williams. One argument in support of his position is Locke's thought-experiment quoted at the start of the chapter. Parfit also asks us to consider this example:

> I enter the Teletransporter. I have been to Mars before, but only by the old method, a space-ship journey taking several weeks. This machine will send me at the speed of light. I merely have to press the green button. Like others, I am nervous. Will it work? I remind myself what I have been told to expect. When I press the button, I shall lose consciousness, and then wake up at what seems a moment later. In fact I shall have been unconscious for about an hour. The scanner here on Earth will destroy my brain and body, while recording the exact states of all my cells. It will then transmit this information by radio. Travelling at the speed of light, the message will take three minutes to reach the Replicator on Mars. This will then create, out of new matter, a brain and body exactly like mine. It will be in this body that I will wake up.
>
> (Parfit 1984: 199)

Parfit's response to the thought-experiment is to confirm that what he has been led to expect is indeed what would happen (1984: 285); but since personal identity is retained despite a total physical change, physical continuity is not even a necessary, let alone sufficient, condition for identity.

I will not attempt to resolve this or any other dispute at this stage, and there are other significant disputes among the analysts which deserve a mention. Perhaps the last major one in the literature which requires some attention is between

adherents of the view that some sort of physical continuity is the relation which constitutes our identity. We have seen Williams's contention, with its supporting thought-experiment, that bodily identity is what constitutes the identity of persons. There is an opposing view that this contention is too strong. While the emphasis is still on the physical, it is argued that it is merely some part of one's body that must continue in order for identity to be retained. That appears to be the consequence of this thought-experiment of Sidney Shoemaker's:

> One day,...a surgeon discovers that an assistant has made a horrible mistake. Two men, a Mr Brown and a Mr Robinson, had been operated on for brain tumours, and brain extractions had been performed on both of them. At the end of the operations, however, the assistant inadvertently but Brown's brain in Robinson's head, and Robinson's brain in Brown's head. One of these men immediately dies, but the other, the one with Robinson's body and Brown's brain, eventually regains consciousness. Let us call the latter "Brownson." Upon regaining consciousness Brownson exhibits great shock and surprise at the appearance of his body...When asked his name he automatically replies "Brown." He recognizes Brown's wife and family (whom Robinson had never met), and is able to describe in detail events in Brown's life, always describing them as events in his own life...Over a period of time he is observed to display all of the personality traits, mannerisms, interests, likes and dislikes, and so on that had previously characterised Brown, and to act and talk in ways completely alien to the old Robinson. (Shoemaker 1963: 23-24)

Here the transfer of Brown's brain into Robinson's body takes with it the identity of Brown, or so it has been claimed[5]. But it is quite obvious that Brownson's whole body, or even most of it, plays no part in that transfer of identity. What has happened to Brown seems to be best described as a radical amputation, and that marks the whole-body criterion of personal identity as misguided.

That provides a rough sketch of at least the major concerns in the personal identity debate. It by no means reflects all the sophistications, nor refers to all of the philosophers who have concerned themselves with the problems. But it does serve to show the kinds of issues which have interested philosophers, and makes quite clear that a singular sort of method has driven the debate. Thought-experiments are by no means to be seen as the preserve of a particular school of thought on the topic, as the wide variety of conflicting views represented shows; everybody seems quick to produce the thought-experiment which proves their point.

SECTION 4: *Adverse reaction to the general method*

The impression that all concerned accept the method of thought-experiment is misleading, for their appeal has <u>not</u> been universal. The various parties in the debate have long accused each other of misinterpreting particular thought-

experiments, or misrepresenting what is at stake, and thus abusing the method in their arguments. But there is also a much stronger and more far-reaching kind of censure to be found in the literature. This ranges from Williams's mild warning that "the imagination is too tricky a thing to provide a reliable road to the comprehension of what is logically possible" (Williams 1966: 45) to Flew's total rejection of the use of thought-experiment (Flew 1988). Wilkes's book Real People is tellingly subtitled, "Personal Identity Without Thought-Experiments" (Wilkes 1988).

Objections to the method range across a wide front; there are complaints of semantic, epistemological and metaphysical problems besetting any attempt to discover the truth about our identity by indulging in imaginative experiments. Nor are these criticisms isolated ones; recently they have increased greatly in number and have received widespread acceptance[6]. The crucial role that thought-experiments have been shown to play, together with the general presumption by their users of their effectiveness, demands that these criticisms receive attention. In the following chapters I will look both at criticisms of specific thought-experiments and more general rejections of the entire method.

This tide of adverse criticism of the use of thought-experiments in the personal identity debate demands attention if one still wishes to make use of them. Thus Unger is obliged to begin his Identity, Consciousness and Value (Unger 1989) with a chapter outlining and defending the method before he gets on to use it to support his view of identity in the

eight chapters which follow. But the problems the method raises cannot be adequately dealt with as quickly as Unger proposes. That is a strong reason for my thesis taking the form it does - almost the inverse of Unger's - of a detailed defence of thought-experiments in the context, followed by an examination of some of the consequences for personal identity.

One point is worth making here, however, even before any detail is provided of the case against thought-experiments. This is that there is a function for thought-experiments even if there is a flaw in the method itself. Even if all the arguments against thought-experiments were to succeed, there would still be a place for them. Once it is accepted they have played, and still play, an influential role in the personal identity debate, their place is assured; for they can work as internal criticisms of the work of theorists who make use of and accept them as a valid method. One is entitled to use a theorist's own thought-experiments, as well as relevantly similar ones which he does not use, against him. The method of thought-experiment would be accepted provisionally and used to undermine the conclusions of those who are satisfied with the method, much as the premises of the non-sceptic might be provisionally accepted in order to demonstrate the truth of scepticism. In doing so, one would not commit oneself to the independent validity of the method[7]. So given the widespread use of thought-experiments in the literature on personal identity, even in the worst scenario, there remains important work for them to do.

SECTION 5: *A general outline of the thesis*

The thesis is divided into two parts. The first part is a defence of the use of thought-experiments against a number of influential and potentially damaging indictments of it. Some of the arguments discussed are directed at specific experiments or a specific kind of experiment, but all have implications which extend to the method in general. The thrust of my response to these arguments is that even if some objections to thought-experiments are strong enough to make us more cautious about how we use them, none of them is strong enough to require the general abandonment of the method of thought-experiment in the context of the personal identity debate.

The second part has a less defensive aim, namely to find an answer to the question of what it is that thought-experiments <u>can</u> do, given that there is no prior case ruling them out altogether. The strategy is to reach an answer by a close examination of some prominent examples of thought-experiments in the literature, although I will by no means always draw the same conclusions from the experiments as their proposers do. In the nature of my topic, there are two issues here. One is methodological, about what one can expect from a thought-experiment; the other is the substantive one as to what thought-experiments can really establish about the nature of personal identity.

This setting-out makes the two parts sound distinct, but of course they are not. In defending the general validity of thought-experiments against the various attacks in the literature, quite a bit will begin to emerge about specific results that thought-experiments are and are not able to achieve. Likewise, in investigating what it is that thought-experiments show in particular cases, more critical points about their use and abuse will come out and have to be discussed. Especially in Part Two, perhaps, the two issues will tend to run into each other. This need not reflect any confusion, for the issues sometimes <u>are</u> intertwined, and I will do my best to signal just how far we have got on each issue at various stages.

Ultimately, the argument of Part Two is that there are two kinds of thought-experiment reflecting two things that thought-experiments can do: there are those which reveal the implicit criteria we use in applying concepts like <u>person</u> and <u>same person</u>[8], and there are revisionary thought-experiments which set out to show, or end up by showing, that there is something wrong with the beliefs or practises which underlie our concepts.

In the process of investigating how thought-experiments can work, I will argue that one view of personal identity receives stronger support from them than any of its rivals. This is a nonreductionist view which holds that while personal identity can be analysed in terms of psychological continuity, it cannot be <u>reduced</u> in Parfit's sense of that term; the relations into which it can be satisfactorily analysed contain

irreducible references to persons. This view and its attendant distinction between analysability and reducibility will be spelt out in more detail in Chapters 7 and 8.

SECTION 6: *A chapter-by-chapter synopsis*

The section headings in each chapter should provide an idea of the strategy of the thesis, but let me conclude this introduction with a more detailed synopsis, outlining what will happen chapter by chapter. Part One, the defence of thought-experiments, consists of Chapters 2 to 5. Chapter 2 takes on an argument presented by Shoemaker and Flew, to the effect that the meanings of the terms "person" and "same person" are grounded in our actual experience, and that this rules against our taking seriously fictional or counterfactual scenarios in investigating the concepts represented by those terms.

Chapter 3, as its title suggests, is concerned with an attack on thought-experiments from an epistemological angle. It is not only epistemology that provides the cutting-edge, however, as semantics once again plays a large part. The central argument to which I respond is one of Fodor's, which suggests that the method of thought-experiment asks us to make knowledge claims for which we can have no reliable warrant. I reject Fodor's argument, but in doing so acknowledge that one

semantic theory which is extremely influential in contemporary analytic philosophy - namely, the causal theory of meaning associated with Kripke and Putnam - appears to provide Fodor's case with strong independent support. I argue that this support is ultimately not forthcoming.

In Chapter 4 I turn attention to an argument of Wiggins which focuses on thought-experiments which try to draw conclusions from the apparent possibility of one person splitting into two. Wiggins's argument, which has consequences beyond this experiment alone, denies that we need to take such cases seriously. Not doing so might make matters easier for all, but even so Wiggins's way out turns out to be unsatisfactory, as does an attempt by Kitcher to refine Wiggins's position.

Wilkes's recent attacks on thought-experiments involving splitting are the subject matter of Chapter 5. While her arguments have something in common with those of Wiggins, she also brings new considerations to bear. My argument in the chapter is that none of these new considerations is any more successful than those Wiggins raises.

Part Two starts its investigation of what thought-experiments <u>can</u> do with a close look in Chapter 6 at Bernard Williams's famous examples in "The Self and the Future". A number of responses to Williams's scenarios are canvassed, and a conclusion about methodology is reached with some, albeit inconclusive, implications for personal identity itself.

Chapter 7 has as its focus Derek Parfit's argument based on his "splitting" thought-experiment. While I have been at

pains to defend such thought-experiments in Chapters 4 and 5, little or nothing has been said up to this point about what the significance is of imaginary cases of splitting. It is pointed out that this case represents a different kind of thought-experiment from those in the previous chapter, a kind which attempts to make us revise some of the beliefs we hold, rather than making implicit ones explicit. It also emerges that it is not at all clear that the conclusions Parfit reaches are the correct ones to draw; in the light of other thought-experiments, a very different picture begins to take shape.

In the discussion of Chapter 7, certain thought-experiments are used which have a focus that is somewhat different from either of the main types examined so far. Their concern is with personal identity <u>across worlds</u>, rather than personal identity over time. While identity across worlds involves different considerations from identity across time, debate about the former informs debate about the latter in important ways. But these new experiments give rise to new objections. I pay some attention to the new objections and then turn to the question of whether similar problems occur with regard to experiments directly concerned with trans-temporal identity. The discussion focuses on an experiment which Parfit calls the "Combined Spectrum", an important one which is designed to show that nonreductionist views of personal identity are false. It is argued that, partly as a result of these new objections, the experiment fails in its aim. However, in the course of the argument more points in

favour of the nonreductionist view of identity which emerged in Chapter 7 come to the fore.

NOTES

1. The references to Locke's Essay are to Book, chapter and section. Thus "II,xxvii,15" refers to Book II, chapter 27, section 15.

2. Indeed, Rescher points out that it starts where Western philosophy starts: with the Pre-socratics (Rescher 1991).

3. Given this remarkable fact, it is fairly surprizing that Sorensen in his book Thought Experiments makes only three direct references to the personal identity debate and thought-experiments found there (Sorensen 1992: 312).

4. As well as for many others, such as the Shoemaker of Shoemaker and Swinburne (1984), and Locke himself.

5. For example by Williams (1970: 47ff). Shoemaker's response to his own example is more complex than, and to some extent at odds with, the story which I tell here (Shoemaker 1963: 24 and 246-7). Nevertheless, Shoemaker does wish to assert a physical criterion of identity. His response is discussed further in Chapter 2.

6. The following list of works expressing opposition gives some idea of how widespread the censure of thought-experiments regarding personal identity has become: Baillie 1990, Collins 1982, Johnston 1987, Kitcher 1979, Lowe 1990, White 1989, Wiggins 1980, Wilkes 1988.

7. Elliot (1991: 58-59) is careful to distance himself from any claim that the method is a generally useful one, and excuses himself on the grounds that he only uses the device against those who use it themselves.

8. Here I underline the terms referring to concepts. I will continue with this practise throughout the thesis, in the hope that it will help me to avoid any use/mention mistakes. Rather than fall into the trap into which Leibniz and others fall (Mates 1989), I signal my procedure at this stage.

PART ONE

IN DEFENCE OF THOUGHT-EXPERIMENTS

CHAPTER 2: BODY-SWAPPING AND THE MEANING OF "PERSON"

SECTION 1: *Body-swapping in the literature*

The modern discussion of personal identity begins with the work of Locke. It is right here at the beginning that the use of thought-experiments begins as well. That provides reason enough to start the discussion of the case against thought-experiments with an argument aimed at the kind of experiment which Locke uses. While the argument has this specific focus, it will become clear that it has general repercussions for the method itself.

The most prominent thought-experiment of which Locke makes use concerns an apparent "body-swap": a scenario in which one person seems to exchange bodies with another. Locke outlines a case in which the soul of a prince "enters and informs" the body of a cobbler (Locke 1694: II,xxvii,15). But we need not stick with Locke's example; a more convenient starting-point - convenient not least because of the absence of talk of souls - occurs in a more up-to-date version of the thought-experiment used by Shoemaker (Shoemaker 1963: 25-28 & 243-245). This version also lends itself to our purposes because it leads into a strong attack on the method by

Shoemaker himself. I will outline the experiment, and then discuss Shoemaker's response to it.

Shoemaker, to recall, asks us to consider the following situation. Brown and Robinson undergo operations whereby their brains are removed from their respective heads. After the removal the brains are replaced, but not in their original heads: the body of Brown receives Robinson's brain and vice-versa. One of the resulting men dies immediately; the other, the one with Robinson's body and Brown's brain survives. He is dubbed "Brownson".

This thought-experiment has frequently been used as an argument to show that it is psychological continuity rather than bodily continuity which is a necessary and sufficient condition for personal identity. The relevant intuition is that the only plausible response is to say that Brownson is identical with Brown. After all, Brownson would have all of Brown's memories, personality traits, projects and so on. The experiment has been taken as showing at the same time that "human" and "person" are distinct kinds, since their respective identity conditions are clearly shown to be different. It has also been used, as in Chapter 1, to argue for a more specific physical criterion of personal identity than a straightforward bodily criterion.

But these are not the conclusions that Shoemaker draws[1]. Granted, the experiment is not used in the context of the issue between bodily and psychological criteria for personal identity or the human/person distinction. Even so, Shoemaker is wary of drawing the conclusions which have seemed to most

commentators to be obvious. He holds that the criteria people currently use to judge personal identity are predominantly physical ones, and that these are adequate for dealing with the situations they encounter. As far as the case of Brownson is concerned, since its kind has never been encountered, there is no established response to be made:

> The question of what most people would say if the imagined events occurred is of course a factual question, and not a question for philosophers to decide. But something can be said, of a philosophical nature, about what would be the case if such events were to happen and if nearly everyone were to agree that a change of body had taken place. First, it clearly cannot be said that in making this judgement people would be mistaken; at most it can be said that in making it they would show that they had adopted new criteria of personal identity and that their judgement would not be in accord with our present criteria. (1963: 246)

Furthermore,

> It cannot be said that they would be abandoning bodily identity as a criterion of identity; all that can be said is that they would be refusing to regard this criterion as decisive in all cases, and would be allowing it to be overweighed by other criteria in some circumstances. (1963: 246-247)

Why is Shoemaker so wary of drawing any strong conclusion from the thought-experiment? If all we can draw from the experiment is that if it were actually to happen and people were actually to agree that the resulting person were Brown,

then they no longer take the bodily criterion of identity to be universal, then the experiment is not nearly so interesting as it at first seemed. He is claiming that what people would say about the identity of Brownson is philosophically irrelevant, and that the fact that we intuitively accept that Brownson is Brown implies nothing about the nature of persons, or about our concepts of person and personal identity. But all this needs explanation, for it certainly goes against traditional thinking on the use and usefulness of such speculations. Nor is it just this particular experiment that is affected: for if our intuitive reactions to a counterfactual situation are philosophically irrelevant, then thought-experiments must to a significant degree be dropped from philosophical methodology. That is a sweeping claim since, apart from their centrality to the personal identity debate, they occur to a lesser degree throughout philosophy.

SECTION 2: *Criticism of the response of Shoemaker and Flew*

What need immediate investigation are the considerations which underlie Shoemaker's response to body-swapping. He is by no means alone in his circumspect treatment of such cases[2]. One of the most forceful statements of relevant considerations, and an influential argument against the use of

thought-experiments to reach conclusions about personal identity, has recently been published by Flew. Flew writes:

> ...we ought never to forget, what almost always in the present context has been forgotten, that there is a categorical difference between fact and fiction. Our notions both of persons and personal identity evolved in adaptation to the actual situations in which our ancestors found themselves; and they will no doubt continue to evolve if and in so far as our actual situations become in relevant ways drastically different. But considerations of how in future we either ideally should or in fact would alter these or other concepts, were we in truth confronted by this or that unquestionably conceivable yet way-out fantastic predicament, are simply not relevant to investigations of the present meanings either of the word 'person' or of the expression 'same person'. (Flew 1988: 123)

This extract expresses precisely similar thoughts to those underlying Shoemaker's discussion, and like Shoemaker, Flew is in effect ruling the use of almost any counterfactual thought-experiment out of court.

Shoemaker's and Flew's wariness stems from the fact that the apparent body-swap described does not actually occur, and the belief that only our responses to actual situations can be an index of the meaning of our words. By contending that consideration of situations which we do not or have not encountered is irrelevant to the meaning of the expressions we use, Flew argues that no interesting counterfactual thought-experiment has any place in philosophical methodology[3]. But

before we curtail our methodology and face the radical consequences of such a move, the case for doing so needs closer attention.

Flew's contention is that the meaning of our terms is exhausted by the actual situations to which they apply. But the argument which he produces - that our concept of personal identity has been formed and has evolved in response to actual situations - far from establishes the irrelevance of thought-experiments. For it is certainly true that however easily we may apply the concept of personal identity in practice, we would find it much more difficult to say precisely why we apply it in any given case. Even though we apply the concept easily, this brings no guarantee that we have any clear idea of the criteria we employ in its application. On the face of things, thought-experiments stand to be able to inform us on exactly this issue, for the following reasons.

There is a common-sense distinction to be drawn between the features which are essential to a kind of individual and features which instances of that kind have only accidentally. By way of illustration, consider a light-bulb. Although all light-bulbs are made of glass, this is merely an accidental and not an essential feature of light-bulbs. In support of this, we can point to the fact that we would not refuse to call an object newly come on the market a light-bulb simply because it was not made of glass. This is the case even though all the light bulbs we have actually come across have been glass ones. So the fact that we do not take some (up until now) universal feature of light-bulbs to be an essential

feature is established by a thought-experiment - that is, by considering how we respond to a certain kind of counterfactual situation - in this case coming across a new kind of light-bulb that is not made of glass. Although the situation is in certain respects an unfamiliar one, we have no hesitation in responding to it. From past practice we know implicitly that being made of glass is not a feature relevant to an object's being a light-bulb, and thus to the meaning of "light-bulb". It is how a thing functions that makes it a light-bulb.

Flew's insistence that the meaning of a term is exhausted by the actual situations to which it applies suggests that our response to the proposed counterfactual situation is irrelevant. But the result of this is to allow no adequate distinction between essential and accidental features. On the basis of actual experience alone, we would have to say that being made of glass was part of the concept of a light-bulb. If counterfactual situations are, as the extract from Flew suggests, irrelevant to the investigation of the meaning of a term, then we have no clear way of distinguishing universal but accidental features from essential ones. Not only is this to deny an intuitively important distinction, but in the process Flew also underestimates the direction that implicit knowledge based on past practice can give to thought-experiments like that above.

As another example, one can consider the concept of a scientist. The context in which that concept emerged was one in which all the instances of the concept were male. We know, however, that masculinity forms no part of the concept of a

scientist. Of course, nowadays we have many female scientists, but the point is that the actual origins of the term did not exhaust its meaning - it did not change meaning once women took up what had been exclusively male practices. Even before then, a thought-experiment could have brought out the implicit knowledge that masculinity was not an essential feature of the scientist.

We can now apply the discussion to the concepts which are our particular concern. Flew and Shoemaker point out that the actual situations in which our concepts person and same person arose involved experience of embodied persons only. They hold that our concepts are accordingly of physical entities and physical continuity. Flew is no doubt right that we don't meet disembodied persons; but it does not follow, despite Flew's apparent inference to the contrary, that our concepts essentially involve embodiment and physical continuity.

Although the context in which the concepts form may be one in which all the persons we meet are humans, and in which persons and bodies don't go swapping around, this does not dictate that in our concept of person we do not distinguish between humans and persons. It does not close off the possibility that embodiment is an accidental and not an essential aspect of personhood. Flew's admission that something like a disembodied person is "unquestionably conceivable" seems to support this point. That is, he seems by this admission to allow precisely what is at issue - that a disembodied individual could be a person, despite our lack of experience of such an individual.

Flew does actually step down from the strong position suggested by the extract above, and he acknowledges that it does not follow from the fact that the only ϕ's experienced are ψ that it is incoherent to suppose that there are ϕ's which are not ψ (1988: 103), but not as far as regards the concepts of interest to us. He thinks that it would be absurd to take ϕ's here as people and ψ as the property of being embodied and thus suggest that persons could exist without bodies. If one were to do so, he claims, one would be unable to re-identify a person or even identify a person as such in the first place. But this hardly follows: just because we usually (or even always) do use physical traits in identifying and re-identifying people, that does not imply that people are <u>necessarily</u> embodied. Flew sets out the following challenge:

> To characterize something as incorporeal is to make an assertion which is at one and the same time both extremely comprehensive and wholly negative. Those proposing to do this surely owe it both to themselves and to others: not only to indicate what positive characteristics might significantly be attributed to their putative incorporeal entities, but also to specify how such entities could, if only in principle, be identified and reidentified.
>
> (Flew 1988: 103)

The implied suggestion is that this cannot be done. But it is by no means clear that Flew is correct in his contention that we would be unable to identify a disembodied person as such or

reidentify them or ascribe positive properties to them. Gillett outlines the following story which strongly suggests otherwise.

> Consider a family called Brown who live in what is supposed to be a perfectly normal semi-detached home. Imagine that things of an unusual nature start happening. Lights go on and off and things are moved in the house. Other things are 'tidied away' or interfered with in unaccountable ways. Each member of the family is suspected but absolved of any blame. One day the father, Mr Brown, conjectures that the house may be haunted by a poltergeist. After he discusses this with the family, to the amusement of some members and the wonderment of others, one of the children begins to receive premonitions of what is going to happen. She says that a person, an invisible person, P, has 'talked' to her. She then qualifies this and says that the person had not really 'talked' to her but rather 'let her know' like 'thoughts popping into her head'. One night she announces that P is going to 'come out of the closet' as it were. She says that P wants to belong to the family and be accepted and that she feels very lonely. Sarah suggests that the others can 'tune in' by adopting the right attitude to P. That evening, at supper, P moves an ashtray across the coffee table, closes the curtain and lights the gas to boil the kettle. The family are amazed. Gradually they learn, over the next weeks, to recognize certain 'thoughts' as being messages from P. P becomes a family friend an lets them know that her name is Polly. P establishes her familial position when Mrs Brown feels a sudden premonition of danger to Sarah, who is playing in the back garden. She runs out and finds Sarah asphyxiating in a plastic bag. From this point on the family become absolutely convinced that Polly is real.
>
> (Gillett 1986: 377-378)

Polly in this story is certainly disembodied and yet she manifests the ability to perceive, to communicate and to act as well as distinct character traits. How she manages to do all this is mysterious, but the story is by no means unintelligible. And given that she manifests these abilities and traits, a denial that she is a person is extremely implausible. Flew's challenge thus seems to receive a straightforward answer.[4]

SECTION 3: *An alternative model to Flew's*

The whole objection to Flew's position can be made in a slightly different way by setting out a more plausible model for the content of a concept like "person". Flew has suggested that thought-experiments involving fictional situations are irrelevant to the investigation of the meaning of "person" and "same person" on the grounds of there being a rigid connection between the meaning of these terms and the actual situations in which we encounter persons. Implicit in his criticism is the view that such actual situations exhaust the meaning of "person" and "same person". In this section I wish to respond to this by outlining an alternative model for the meaning of the relevant terms. I will argue that the model is at least as plausible as Flew's, and yet supports a distinction between essential and accidental features and thus

undermines the rigid link between meaning and actual situations which Flew sees as so important. I will also argue that not only does this alternative model allow thought-experiments to be relevant, but that the ease with which we respond to thought-experiments and other fictional situations provides a form of evidence for the truth of the model.

The alternative model to Flew's stems from David Lewis's account of the meaning of mental terms (Lewis 1972). All of the most influential accounts of the concept of a person stress the centrality of certain mental concepts. For example, Locke defined a person as a "thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places" (Locke 1694: 39). Daniel Dennett's widely acknowledged "conditions of personhood" include that to count as a person an individual must at least be rational, self-conscious and be the subject of intentional attitudes (Dennett 1976: 269-270)[5]. These examples serve to illustrate the general acceptance of a crucial link between mental concepts - especially those of "intentional attitudes" like belief, desire and intention - and the concept of a person. It is here that Lewis's account of mental terms becomes relevant: there appears to be a strong connection between the meaning of "person" and the meanings of our mental terms, and Lewis offers us an account of those. What is important for our purposes is that that account is consistent with the distinction between essential and accidental properties discussed above.

Lewis suggests that our mental terms are "theoretical terms". That is, they are terms which get their meaning from the theory in which they occur. In this case the theory is our folk-psychology, the theory made up of commonsense principles according to which we explain and predict the behaviour of those around us. A term will be implicitly defined by the role it plays in the principles of this theory.

Take, for example, the term "desire". Following Lewis's model, the term is defined by the role it plays in folk-psychology. The role it plays there is a causal or functional one. The relevant principles which serve to outline its role in our system for explaining behaviour are epitomised by the practical syllogism: if X desires that p and believes that doing A will lead to p, then X tends to do A. Desire is the state which typically causes one to act to get something, and which is typically caused by some or other lack, or by a perceived need. In the same way the meaning of our other mental terms can be spelt out by reference to the causal role they play in other parts of this explanatory system.

Using this brief exposition of the alternative model to Flew's, it can be made clear how Flew's objection to thought-experiments can be undermined. The mental terms which are of interest to us are defined by the roles they play in folk psychology. Thus belief, desire and other mental states are defined according to the functional roles they play. Since the principles of folk psychology define desire (to use the same example) as a functional state of this kind, what emerges is that other features of actual desires - features not linked

in any fundamental way to the typical causal role of those desires - are not essential to those states being desires. So, for instance, although our actual desires are associated with our organic brains this need not have been the case. Having an organic brain is not essential for having a state with the causal role typical of desire. Even if all individuals which actually have desires also have such brains, this is beside the point - being associated with that kind of brain is an accidental feature of our desires.

This all has relevance for Flew, because the definitions of mental terms implicit in folk-psychology will set boundaries for the type of situation in which the terms will apply, types of situation which may be found instantiated in otherwise bizarre circumstances which nobody has actually experienced. As long as our mental concepts like that of desire are applied according to implicit principles like those indicated above which underlie our explanations of behaviour, then we will be able to distinguish between essential and non-essential conditions for the application of those concepts. This means also that we will be able to apply the relevant terms in at least some counterfactual situations. As long as the essential conditions for the concept's application are satisfied, then we will be able to use the concept, no matter how grossly the situation considered differs in other ways from situations we have actually experienced. Just as the principles of science can be applied in counterfactual situations, so can the principles which (on the model under consideration) go to define our mental concepts. And as a

result the distinction between fact and fiction is not nearly so categorical as Flew suggests.

I have maintained that Lewis's model is a plausible one, and as such it is enough to undermine the arguments of Flew and anyone else who believes that our concepts must fail in strange counterfactual situations. Wittgenstein has been taken to hold a position similar to that of Flew in passages such as this one[6]:

> If you imagine certain facts otherwise, describe them otherwise, than the way they are, then you can no longer imagine the application of certain concepts, because the rules for their application have no analogue in the new circumstances.
> (Wittgenstein 1967: §350)

As this claim stands it is far too sweeping, just as Flew's claims are. The application of our concepts will only be unimaginable if the facts which are changed are such that the essential conditions for that concept will no longer be satisfied. But Wittgenstein can also be read as making this narrower claim, and the existence of this plausible alternative to Flew's model suffices to show Flew's opposition to thought-experiments to be unreasonable.

Although that is sufficient to make the case, there is more to be gained from this discussion. For the very fact that our concepts do seem to apply in certain strange

circumstances but not others provides evidence of a sort for the view that our mental concepts work according to implicit principles along the lines suggested above.

Perhaps the most striking example to consider is that provided by cartoons and the characters that appear in them. In cartoons we discover characters like Bugs Bunny who, although he is a rabbit, has sophisticated beliefs and desires far beyond any we might encounter in actual rabbits. Indeed, we would only find humans having mental states as complex as the desire to discover the weak points of one's enemy (in this case, Elmer Fudd) and exploit them to the full. We cannot reasonably assert that Bugs Bunny is a human, yet our concept of desire fits neatly in Bugs's cartoon world. In other words, this implies that you do not have to be human to have extremely sophisticated desires, even though it is only humans who we actually experience as having such states. Nor does this only apply to animals like Bugs Bunny. In Walt Disney's films we meet cellos and teapots to whom nobody has any trouble ascribing beliefs and desires. This all suggests that there are implicit principles according to which these concepts are applied, and which rule being human or being organic as non-essential conditions for their application.

A more detailed example may serve to make the point more clearly. Consider the case of someone who suffers a brain trauma and as a consequence loses motivation. Once the problem is diagnosed, the person undergoes micro-surgery in which the affected parts of their brain are replaced by electronic components designed to perform the tasks once

performed by the damaged brain cells. Organic brain cells have now been replaced by functionally equivalent non-organic parts. After the operation, the person's behaviour patterns return to what they were before the brain trauma. Intuitively it is appropriate to describe the person as having recovered the capacity for desire. That is, we would ascribe desires to this person even though those desires are not organically realised as our desires happen to be.

It is not the case, then, that our concepts must fail in counterfactual contexts. "Way-out fantastic predicaments" are only irrelevant to the meanings of our words if they preclude the essential conditions for the application of the relevant concepts. Were we to come across creatures who were totally self-contained and self-supporting, needing nothing from their environment, our concept of desire would have no purchase where they are concerned. Our reaction to this situation confirms that the essential conditions for applying the concept of desire are not satisfied in the case of creatures that never act to get anything, and provides further confirmation for the model proposed.

Just as counterfactual contexts can be relevant to mental terms like "desire" and "belief", so they can be to terms like "person" which these mental terms go to define. Thus just because persons have not changed bodies in our actual experience, it does not follow that imagining cases in which they do will result in examples where our existing conceptual apparatus - and in particular our concept of same person - has no application. And as a result, Flew has not presented us

with a case which should lead us to reject the use of thought-experiments as a tool in the personal identity debate.

NOTES

1. It should be noted that Shoemaker has revised his views since the early work cited, and in Shoemaker (1970) and Shoemaker and Swinburne (1984) he draws conclusions very much like those mentioned.

2. Quine also voices a similar objection. He contends that

> to seek what is 'logically required' for sameness of person under unprecedented circumstances is to suggest that words have some logical force beyond what our past needs have invested them with.
> (Quine 1972: 490)

This amounts to much the same objection to those of Flew and Shoemaker, and I will just deal with the (fuller) position of Flew explicitly.

3. No thought experiment has a place, that is, apart from his own special use of a thought experiment in Flew (1951).

4. To argue in this way is, of course, to use a thought-experiment; but they can presumably be regarded as innocent until proven guilty.

5. As Dennett points out (1976: 269) these conditions are common themes in all accounts of personhood.

6. Parfit takes Wittgenstein as holding this sort of view (Parfit 1984: 200).

CHAPTER 3: KNOWING WHAT WE WOULD SAY

SECTION 1: *Fodor's argument that we cannot know what we would say*

One influential line of argument against the use of thought-experiments concerns an important epistemological question. Broadly, the argument runs something like this. The point of outlining some counterfactual scenario is to ascertain what we would say in the situation described. Knowing this will give us insight into the necessary and sufficient conditions for applying a concept; or into whether F is an essential feature of G's rather than an empirically adequate criterion for claiming something to be a G. Thus we are called on to imagine a situation in which some x seems to be a G, but is not an F (for example), and are asked what we would say - would we say x was a G or not?

So the point of thought-experiments is to discover the present meaning of our words by finding out what we would say in some crucially different situation. The problem, Fodor argues, is that we cannot know what we would say in situations which are crucially unlike our actual ones.

This general line of argument occurs in an early paper of Fodor's (Fodor 1964), and it comes up again in Wilkes's recent

book, Real People (Wilkes 1988). For the purposes of this chapter, I will concentrate on Fodor's argument, looking at Wilkes's slightly different case in Chapter 5.

The brief exposition above misrepresents Fodor's position slightly. For rather than making the strong epistemological claim that knowledge of what we would say in unprecedented circumstances is impossible, he argues that it would be irresponsible to rely on our intuitions in this matter. Let me spell the argument out in more detail.

Fodor's first point is to stress that the thought-experiments under discussion involve counterfactual situations. On top of this they are counterfactual situations such that we have no way of testing observationally whether what a speaker claims he would say is what he really would say. The situations are of interest precisely because they are abnormal: certain features which usually obtain do not obtain in them, or certain features obtain in them which do not obtain in actual experience. The point of this is to get to the features of some word or concept which are determined by meaning alone; features which are essential to a concept, and which are not, though universal, merely accidental accompaniments to instances of that concept.

As a result, the strategy is to reach a position where we can rely on a speaker's mastery of his language: on his intuitions as to whether or not a given concept still applies. It is here that the problem arises. For the methodology requires that we rely on a speaker's present linguistic intuitions, and yet there is much more than this operating

here. We are being asked to say what we would believe were certain usually reliable features of our experience to cease to be the case. For example, one is asked to contemplate a scenario in which people apparently swap bodies, or divide into two. But in supposing these things one is supposing that certain laws of nature, or at least certain true empirical generalizations, no longer hold. Thus we are being asked to say what beliefs and what theories we would adopt should our current beliefs or theories prove to be false.

Fodor contends that in most interesting cases this is just asking too much. For how can we possibly know what theories would come to be accepted should our present ones be rejected? For instance, how do we know what people would come to believe were all the water on earth to turn red? How many of our current beliefs would we give up? How would other theories actually be affected? We certainly have no general principles according to which such a judgement can be made; knowing what we would come to believe is not just a matter of working out what claims are implied by the claim that water is red, building those into our scheme and rejecting beliefs incompatible with these. We just have no vaguely reliable way of predicting how beliefs and theories would evolve should something like this happen. More to our point, it would be just as, or even more, difficult to predict how our belief-system would change should people suddenly start swapping bodies or dividing into two on a large scale. How things would pan out is something we just don't know.

There is still a further point to Fodor's criticism. He writes:

> it is unreasonable to attempt to predict what theories would be accepted if our current theories were abandoned and, a fortiori, it is unreasonable to attempt to make such predictions on the basis of an appeal to our current linguistic intuitions.
> (Fodor 1964: 207)

The point stressed here is that the intuitions the thought-experiment strategy requires us to rely on are linguistic ones; intuitions about the meanings of terms which form part of our mastery of language. It is as speakers, and as speakers alone - not as armchair psychologists or sociologists - that we would have to make the prediction as to our beliefs in the situation outlined. But the kinds of beliefs we are being asked to make judgements about - that is beliefs about the world in general - go far beyond judgements which have only to do with language; and so, as speakers alone, we are simply not competent to make the required decisions.

SECTION 2: *Why Fodor's argument misses its mark*

Fodor may well be correct that it would be irresponsible of us to place any great weight on our current ideas about how our beliefs in general would change should some reliable feature of our experience cease to hold. Furthermore, to ask what one would say in such a situation does seem to rely on ideas of that sort. But even if he is right about all of this, his argument is wide of the mark and certainly does not show that we should give up the thought-experiment strategy. The reason for this is, as I will argue, that his argument turns on a crucial misdescription of the strategy of thought-experimentation. He may be correct that some philosophers have used the strategy he describes and attacks, but if so, this shows only that they are guilty of misusing the method, and not that the method itself is faulty.

Fodor's argument requires that speakers be asked to say how they would respond <u>if they were in the situation described</u>. It is because we don't know how they would respond that he takes the strategy to fail. But we would not have to know what we would say <u>in</u> the situation in order to carry out the task of separating the essential from the non-essential conditions for the application of a concept. It is our response, given our present belief-system which matters: we are being asked, "does the concept <u>as it stands</u> allow such-and-such?". Perhaps the point can be made clearer in the following terms. Fodor is claiming that what the thought-experimenter asks is, "What would our language-game <u>be</u> if

such-and-such were the case?". But the correct, and unproblematic, question to ask is, "What would we say in our language-game if such-and-such were the case?"

Take the case of persons swapping bodies. In normal circumstances people do not swap bodies; perhaps such a thing is empirically impossible. That, of course, does not stand in the way of the strategy outlined above: considering a hypothetical body-swap is just the sort of thing to test whether sameness of body is a logically characteristic feature of personal identity or merely an empirically adequate one according to our existing conceptual structures. So our thought-experimenter sets up a scenario in which human body A takes on the psychological characteristics which we used to associate with human B, and he asks, "is the A-body person A or B?". Our intuitions, as did Locke's 300 years ago, suggest that the A-body person is now B; we conclude that sameness of body is not a necessary condition for applying the concept same person.

Now we do not know how our belief-system would change should this sort of thing start happenning, and we cannot predict on the basis of our linguistic intuitions how people would be treated should an operation to bring this situation about become feasible. Perhaps we would all agree to carry on treating the A-body person as A. Who knows? But that seems beside the point. In reaching the response that persons A and B would have swapped bodies we do not, nor do we need to, involve ourselves in claims about how our beliefs would change were we faced by the phenomenon. All we commit ourselves to

is the belief that nothing in our intuitive understanding of what is crucial to personal identity is inconsistent with persons swapping bodies. In Fodor's terms, our intuitions about the meaning of "same person" lead us to say that the A-body person is B, and it seems on the face of things as if our present mastery of the term is all that has been relied upon[1].

SECTION 3: *Independent support for Fodor from the causal theory of reference*

Nevertheless, there is an important issue here which still needs to be discussed. One of the central thrusts of Fodor's argument is that thought-experiments are not capable of revealing semantic facts. In our case the claim would be that they reveal nothing about the semantics of "person" or "same person". Now, while Fodor is wrong about why thought-experiments might fail in this task, there is some independent reason to believe that his conclusion is correct after all.

The independent reason comes from the work of Kripke and (especially) Putnam on the semantics of general terms, "kind" terms, of which "person" is one (Putnam 1975, Kripke 1980). A brief look at their view on such terms will begin to show why Fodor's claim has some plausibility.

The Kripke-Putnam view arises as a response to another influential view on the meaning of general terms. This is the theory due to Frege, Searle and others[2] that the sense of a

general term, like the sense of a proper name, is given by one or more descriptions which we associate with that term. These descriptions will serve to pick out one and only one kind of individual, and in this way sense will determine reference. Crucial to this theory is that in order to use a general term meaningfully - in order, that is, actually to refer to a particular kind of individual - one must know some such identifying description/s.

Against this, Kripke and Putnam have contended that one can perform the task of referring to kinds of thing perfectly well without this sort of knowledge. Indeed, one could still talk about (say) dahlias, even though the beliefs one had about dahlias - or certainly any beliefs one associated with the term which might individuate one particular kind of plant - were totally false. My ignorance of what makes dahlias dahlias does not prevent me from talking about them succesfully. What this sort of example shows is that one does not need knowledge of the facts which individuate a kind in order to refer to its members. It also shows that the descriptions which we do associate with a kind are no more than contingent marks, and are not logically binding as the description theory suggests.

What the success of one's reference depends upon is the existence of a causal link between my use of the term "dahlia" and actual dahlias. At some stage in the past, the word came to be used to pick out this particular kind of plant, and the usage became general. As long as a causal chain can (in principle) be traced between my usage and such a "baptism",

and as long as there is a scientific theory which groups a set of individuals around that original exemplar, then I succeed in referring to dahlias when I use the term.

It is important to note that this account of the meaning of general terms was designed for natural-kind terms like "gold", "tiger", or my example of "dahlia". But if we take its central notions a bit further, then there may well be important consequences for the kind of thought-experiments that concern us. If we assume for the moment that this general picture works for "person" as it does for "dahlia" or "gold", some consequences immediately become apparent. The meaning of "person" and the associated meaning of "same person" would not depend in any fundamental way upon the descriptions we might associate with the terms. In other words, what these terms really mean is not determined by the beliefs we have about the kinds of individuals (or relations) to which they purport to refer, or by the rule-of-thumb criteria we commonly use for applying the terms.

At this stage we can return to the matter at hand: to thought-experiments and the question of whether they inform us on semantic issues relevant to "person" and "same person". It would be of use to have an example to structure the discussion. For this we can use Locke's proposed case of an apparent body-swap.

> should the soul of a prince, carrying with it the consciousness of the prince's past life enter and

> inform the body of a cobbler, as soon as deserted by his own soul, everyone sees he would be the same person with the prince, accountable only for the prince's actions: but who would say it was the same man? (Locke: II,xxvii,15)

Locke is in effect asking us what we would say about the situation which results, and taking it to be the case that everyone would agree that the prince - that person - has swapped bodies. He takes it to follow that personal identity does not co-incide with human or bodily identity.

Using this as our illustration, we could say that Fodor's argument concludes that the thought-experiment tells us nothing about the meaning of the crucial terms: it does not show that bodily identity is not a logically characteristic feature of the term "same person". Fodor's reasons for this conclusion would be those outlined before: that the strategy of thought-experiment is to rely solely on a speaker's mastery of language, that is, on his linguistic intuitions; but that to answer the question as to "what we would say" in a counterfactual situation requires more than this - especially, it requires knowing how the rest of our beliefs will change should certain reliable features of experience cease to hold.

I disagreed with Fodor's argument on the grounds that he conflates what we would say given <u>our</u> conceptual scheme with what our future conceptual scheme would be should the world suddenly become as we have imagined it to be. But reasons other than his for reaching his conclusion now begin to

emerge. Consider what the thought-experiment does. In asking what we would say about the identity of the person in the cobbler's body, it requires us to weigh certain of our implicit principles against each other: namely, the principle that "consciousness of past life" is crucial to personal identity and the principle that being the same human (i.e. bodily continuity) is crucial to identity. In the normal unreflective course of our lives, we are not called upon to weigh these two principles against each other because the memory of a particular life and bodily continuity coincide in the case of all the individuals we know. What the counterfactual situation represented in the thought-experiment does is to show that they need not coincide, and that when they are in conflict we are more strongly committed to the one than to the other. Faced with the hypothetical situation, it becomes clear that we are more easily prepared to give up the principle that bodily identity is crucial than we are its rival.

Have we then learned any semantic facts from the experiment? Not according to the causal theory of reference which I outlined above. For what we have learned concerns which of two of our implicit principles concerning what is crucial to personal identity we take to be more fundamental. This needs to be stressed: we learn something about <u>the principles implicit in our applying a concept</u> and their relative weighting in our conceptual system. As we have seen, the causal theory of reference removes any beliefs or principles in our head from a position of semantic importance;

what we believe to be essential to persons and their identity does not play any crucial role in the meaning of those terms and is no more than a rule-of-thumb. It is only if one accepts a version of the description theory of reference that one could take the thought-experiment to show something about the meaning of the terms concerned[3], but that theory is mistaken. And because of this, Fodor may appear to be right that we do not learn semantic facts from these experiments.

SECTION 4: *Why this support does not help*

I am sympathetic to much of what Kripke and Putnam have to say about the meaning of natural-kind terms. I want to explain why I think those claims do not show the invalidity of thought-experiments in our context. I noted above that the causal theory as outlined is set up specifically for natural-kind terms, and the question must now be faced as to whether this causal model works in the case of "person" as it does for natural-kind terms like those mentioned above. The question is important because there are reasons for believing that it does not.

One reason comes from Putnam himself. In Meaning and the Moral Sciences (Putnam 1978) Putnam despairs of his model working even for the kind term "human being", and the reasons he gives are directly relevant to this discussion. One

feature of the account outlined above is that the reference of a kind term depends on the existence of a scientific theory the laws of which serve to group objects around an ostended exemplar or paradigm. According to Putnam, however, we cannot realistically expect to get a scientific theory of human beings, at least not for a very long time (Putnam 1978: 62). He reasons that having an explanatory theory of human beings would require all our current social institutions to have changed so much as to be unrecognizable:

> would it be possible to love someone, if we could actually carry out calculations, of the form: 'If I say X, the probability is 15 per cent she will react in manner Y'? Would it be possible to have friendships or hostilities? Would it even be possible to think of oneself as a person?
> (Putnam 1978: 63)

Putnam's despair here is rather puzzling, but even so it raises problems for our enterprise. The puzzlement stems from the fact that the considerations he outlines do not really seem to be problems in the way of achieving a theory of human being as a natural kind. Humans are animals just like any others, and there are perfectly adequate biological theories available: theories which will provide accurate enough inclusion conditions for the class of human beings. The kinds of considerations he raises seem rather to affect the

availability of a suitable scientific theory for the kind person.

Here we need to acknowledge that persons do not form a natural kind. As the discussion of Chapter 2 section 3 brought out, it is what a thing does and how it is treated that makes it a person, rather than any matter of internal structure which typically makes biological kinds the kinds that they are. Locke was indeed on to something when he suggested that "person" is a forensic term (Locke 1694: II,xxvii,26) and Putnam is thus right to despair of applying his model as is to such individuals and the term which picks them out. Should we follow his model and look for a theory which groups individuals around a paradigm person, we would end up with the wrong group: that of humans rather than persons.

This point brings us close to the heart of the problem with appealing to the causal theory to back up Fodor's case against thought-experiments, and it will bear closer attention. Human beings are the clearest examples we have of persons. In the hypothetical baptism of a kind of individual with the title "person", then, the original individual around which some theory is to collect a set of similar individuals would most probably be a human being. And yet "person" does not mean the same thing as "human being" and it is at least highly contentious to hold that the extensions of the two terms co-incide. This requires that some modification be made to the account of how kind terms like "person" get their reference.

What needs to be added is an explanation of what it is about the human being picked out in the baptism situation which determines that it is as a person rather than as a human being that the individual in question is being picked out. The only plausible explanation here is that it is the attitudes of the baptizer which determine what exemplar is selected. The baptizer has in mind the fact that a certain individual instantiates certain properties and in virtue of these he dubs the individual a person.

This sort of modification to the causal theory is not only required by terms like "person"; it is even required when it comes to natural kinds. For instance, it needs to be explained what determines that an individual is picked out as a dahlia rather than as an example of one of the other kinds to which it belongs: flower, plant, etc. In the case of the dahlia, it will be various surface properties of the plant which the baptizer has in mind that determine which underlying nature is relevant to the extension of the term applied. These will determine which scientific theory grouping other individuals around that one is relevant (Devitt & Sterelny 1987: 72-75). In both the case of the dahlia and that of a person a change has to be made to the purely causal theory outlined.

In Devitt and Sterelny's terms, a "descriptive-causal" theory must replace the causal one. Certain descriptions which we associate with the term under discussion will play a crucial role in determining its reference. This is not a return to the pure description theory, however. Most

importantly, reference still depends on the term's being causally grounded in some exemplar, and thus one is not required to have knowledge of the crucial descriptions in order to use the term successfully. One may be wrong about what it is that something must be able to do to be a person and yet one would still be able to talk about persons.

It is crucial to notice the undermining effect this move from a causal theory to a descriptive-causal theory stands to have on the mooted support for Fodor's conclusion. What the supporting argument used the causal theory to do was to make the implicit principles we might use in applying the concept of a person irrelevant to the meaning of "person". The argument just presented shows that this cannot be done: if we are to distinguish as we do between humans and persons, then speakers' implicit beliefs play a crucial role in the meaning of "person". I argued above that what thought-experiments like Locke's do is bring us to a realization that there is conflict between certain of our implicit principles regarding the application of the concepts of a person, and to show us which are the fundamental ones; as a result they are at least potential ways to reaching semantic facts.

Although that serves to make the case against an appeal to the causal theory to support Fodor's conclusion, much remains to be said about the reference of "person". I've argued that a "pure" causal theory - one which makes no appeal to associated descriptions or the attitudes of the speaker - is insufficient to explain the semantics of this term. But that does not mean that it is on the wrong track altogether.

In the next chapter I will discuss, and to an extent endorse, an attempt to show how the theory can be modified to provide the required explanation. As will become clear, there are points of significance to our discussion still to emerge.

The conclusion that we should then draw from this discussion is that the causal theory cannot plausibly be used as was suggested to bring support to Fodor's case against thought-experiments. Even if the use of thought-experiments does in the end constitute a problematic methodology in the context of questions about personal identity, they cannot be easily ruled out as irrelevant to the semantics of the case.

NOTES

1. Fodor is not alone in making this mistake. Wilkes, for example gives the following description of the method of thought-experiment.

> ...we imagine a 'possible world' in which (an imaginary) state of affairs actually occurs - a world like our own in all relevant respects except for the existence in that world of the imagined phenomenon...Then we try to draw out the implications - 'what we would say if' that imagined set-up were to obtain; that is, if we inhabited that possible world.
> (Wilkes 1988:2 - my italics)

2. One might wish to include Russell in this list, despite his disagreements with Frege. He certainly also suggests that terms manage to refer via descriptions which we associate with them.

3. This claim that causal theorists cannot use thought-experiments to reveal the meaning of terms may seem surprising in the light of some of Kripke's work. For in Naming and Necessity (Kripke 1980), Kripke concludes on the basis of his intuitions regarding the concept "pain" that what the term "pain" refers to is not a physical state, but a phenomenal quality. Whether or not this squares with his causal account of reference, it serves to strengthen my claim that causal theorists in the end offer no support to Fodor's conclusion that thought-experiments reveal no semantic facts.

CHAPTER 4: SHOULD WE TOLERATE PEOPLE WHO SPLIT?

SECTION 1: *The strategy of the chapter*

The idea that it is possible for persons to divide has long been of interest to philosophers concerned with personal identity. Thought-experiments outlining some scenario in which one person splits into two have been claimed to hold dramatic consequences for the concept of "same person". Thus, for instance, Parfit has used them to argue that personal identity turns out not to be the deeply important notion we intuitively take it to be (Parfit 1984: sections 89-90), and Wiggins has suggested that these experiments (were we to accept them as valid) would threaten the very coherence of the notion (Wiggins 1976, 1980).

Others have been less impressed by these "fission" thought-experiments[1]. But the onus remains on those opposed to show what is wrong with the conclusions drawn from, or with the description and strategy of, the speculative experiments. I believe that this is a more difficult task than is often realized. In this chapter and the next one I wish to look at three important attempts to undermine thought experiments involving the division of persons, attempts which also have

consequences for thought-experiments in general. I will argue that all three fail, leaving us to face the consequences of splitting, whatever they may be.

In Section 2 of this chapter two examples of the kind of thought-experiment in question are set out. In Section 3 an argument against such experiments from Wiggins is set out and criticized. The argument examined in Section 4 is an attempt by Kitcher to solve the problems encountered by Wiggins's argument. In the final section, I will return to some of the points made in chapter 3 and tie up some loose ends.

SECTION 2: *Fission thought-experiments and what they have been claimed to show*

Before looking at the various attacks on fission thought-experiments, it would be useful to have a brief account of the sort of experiment that is at stake. Experiments that appear in the literature do not usually describe one human splitting spontaneously, like an amoeba, into two. More representative is Wiggins's scenario (Wiggins 1967: 52-55) in which one individual's brain is divided and the two halves placed in distinct bodies. Perhaps the most influential example is that put forward by Parfit, a case I outlined briefly in Chapter 1. To recall, Parfit assumes that he is one of a set of identical triplets, and that each of his brain hemispheres is capable of

the functions of the other; he then outlines the following scenario:

> My body is fatally injured as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me. And he has a body that is very like mine. (Parfit 1984: 254-255)

What do these scenarios show? I will attempt to answer this question in Chapter 7. My present aim, rather, is to argue (against various attacks) for the validity of fission thought-experiments. It would be useful, however, to have some notion of what the experiments are claimed by their proponents to show.

Wiggins points out that fission cases stand to undermine standardly accepted criteria for personal identity, and thus the concept itself. In the cases outlined, criteria based on physical or psychological continuity seem unable to ground a decision as to which of the resulting people is identical with the original. Parfit also draws attention to the impossibility of a grounded decision as to identity (1984: 253-265). It would, he argues, be strongly counter-intuitive to say that the original person was dead; that is, that nobody is identical to the original. On the other hand, the logic of

identity prevents both candidates from being the original. But to say that one of the candidates retains the identity of the original while the other is merely similar is to make an arbitrary decision. What all this leads up to, argues Parfit, is that questions of identity can have indeterminate answers, and any theory which implies that they cannot must be wrong. So nonreductionist views of personal identity which will have this implication are shown to be misguided, and personal identity is exposed as not being the deeply important notion it has been taken to be.

SECTION 3: *Wiggins's rejection of splitting persons*

The first argument against fission that I wish to examine is due to David Wiggins. Wiggins's strategy is to argue that "person" is a natural kind term, or at least has an important natural kind element: that the concept of a person is a concept "akin to a natural kind concept" (1980: 172). Although we came across some reason to doubt this in Chapter 3, for the sake of the present argument we can temporarily set these doubts aside; and we will meet some important modifications to the theory discussed there which will make this suspension of doubt worth the while.

The advantages of person being treated as a natural kind like frog or human being are great, the most important one

being that we would not need to be concerned by troublesome cases of fission and the like[2]. To put it briefly, this is because the extension of the kind would be something fairly close to the extension of human being and the natural laws governing humans do not allow fission.

The strategy has its base in Putnam's account of natural kinds and natural kind terms. To recall: the reference of a natural kind term is established by the ostension of paradigms or exemplars of the kind in question, and the extension of the term is the set of objects grouped together with the exemplars by the most comprehensive scientific theory available. Whatever theory fits this description, it will consist in part of a set of natural laws, laws which determine what counts as a member of the natural kind. This is a crucial point in Wiggins's argument:

> Unless there are such laws, the putative kind name has no extension, nor even the sense it is required to have. If there are such laws, on the other hand, then their holding is nothing less than constitutive of the existence of (individuals of that kind). It follows that, <u>if</u> person is a natural kind, then when we consider the problem of the identity of persons through change, the whole logic of the situation must exempt us from taking into account any but the class of situations which conform to the actual laws of the actual world. For these serve...to define the class of persons. (Wiggins 1976: 158)

According to Wiggins, then, our natural-kind concepts have their basis in what is actually the case. The exemplars serving to fix the reference of "person" would be actual human beings and, as pointed out above, the laws governing humans in the actual world preclude any sort of splitting, for in our world humans cannot divide. It would thus be impossible for an instance of the concept person to do so, since on this view what is possible for such an instance is limited by the natural laws which apply to items of the kind in question.

There is, of course, another important issue lurking here. For certain influential thought experiments involving fission are set up as cases which <u>are</u> in accordance with the laws of nature. One example is Parfit's case set out in Section 1. Parfit insists that the division of his brain and the subsequent implanting of the two halves into the brainless bodies of his identical siblings is "merely technically impossible". By this he means that it is empirically possible - that is, it satisfies the laws of nature. It's only that technology is not yet advanced enough for the experiment to be performed; there is nothing "deeply" impossible about separating brain hemispheres, one hemisphere performing the tasks of the other as well, or dividing one stream of consciousness into two (Parfit 1984: 245-246). But if Parfit is right about all this, then it seems that we don't yet have a reason for ignoring speculations involving fission.

Although Wiggins does not address this point in any detail, he would presumably see things otherwise. Although no law of nature is overtly violated in Parfit's description of

his division, there are certain laws with which the phenomenon described may not be consistent, and to which Wiggins draws attention. For among the laws which determine the extension of a natural kind, f, must be those which "define the characteristic development and typical history of individual fs" (Wiggins 1976: 158). The problem with fission is that it represents a vast departure from the typical history of a person; a departure so vast that we are justified in viewing it as irrelevant to questions about the kind person.

SECTION 4: *A response to Wiggins's rejection of splitting*

In this way, treating "person" as a natural kind term avoids any threat from science-fiction thought-experiments. If Wiggins is correct, the envisaged experiments would be impossible in a far more damaging sense than the "mere technical" sense contemplated by Parfit. But a response along these lines is unhelpful when it comes to Parfit's case. This is because it is by no means clear that the thought-experiment is not consistent with the actual normal development of persons. In Parfit's defence it can be held that should our technology advance individuals of otherwise normal personal development could be split. Fifty years ago an argument relevantly similar to Wiggins's could have been produced in support of the claim that persons cannot have (non-human)

animal hearts: receiving one would represent a vast diversion from the typical history of a person. But today baboon hearts have been transplanted into normal humans.

Nothing in Wiggins's account amounts to an argument against this; nothing suggests that normal humans could not be split, or that Parfit's transplant is crucially unlike the baboon-heart transplant. To respond that humans don't actually split is simply irrelevant, and to argue that split individuals would cease to be persons would be absurd. Of course, Parfit could be wrong in his claim that the reasons why splitting is not yet actual are merely technological ones: but that does not help this argument of Wiggins's - for establishing that requires a different kind of argument altogether.

What Wiggins needs is an empirical argument, for the issue of whether humans can split is an empirical one, just as was the issue about humans and animal hearts. There are empirical considerations which might offer Wiggins's anti-splitting stance some support. For instance, the fact that different psychological functions are located in different lobes of the brain would seem to be a problem for Parfit's premise that each of his lobes (in the thought-experiment world) are capable of the functions of the other. But more would be needed even than this, because Parfit's pointing to actual stroke-sufferers who have regained the functions they lost through the loss of the use of one brain hemisphere seems to establish the possibility he requires, even though we will never meet an individual like the one he describes. Parfit

does acknowledge that "it seems likely that it would never be possible to divide the lower brain, in a way that did not impair its functioning" (Parfit 1984: 255). But even if this is true, we still need an argument from Wiggins as to why this impossibility would be crucial in preventing a split. Would, for example, leaving Derek's brothers' lower brains intact and transplanting one of Derek's upper hemispheres on to each make a crucial difference? All this needs supporting evidence if Wiggins's case is to be at all convincing.

Even apart from these considerations, Wiggins realises that things are not quite so easy as his argument presented above might make them appear. Wiggins acknowledges along with philosophical tradition and my arguments of Chapter 3 that, despite the advantages of treating it as such, "person" is not just a natural kind term like "human being". Following Putnam's account of the way the reference of a natural kind term is fixed, pointing out exemplars of the kind <u>person</u> would result in the extension of "person" being the same as that of "human being", and as Wiggins realises (1976: 160), the two terms are not interchangeable. Any account of persons must bring into consideration their character as <u>agents</u> with beliefs, intentions, and so on; something which is not essential in an account of humans. In other words, following Putnam's account as it stands, the wrong class of individuals would be isolated.

Wiggins suggests that this problem can be avoided by adding a "functional specification", setting out the peculiar

characteristics of personhood to a natural kind requirement. Thus

> x is a person if x is an animal falling under the extension of some natural kind whose members perceive, feel, remember, imagine, desire, make projects, move themselves at will and carry out projects... (1976: 161, cf 1980: 171)

This move increases the extension of person so as to include not just humans, but those creatures "who come near enough to us" (1976: 161). The ultimate outcome is that

> every person would belong to some natural kind that determined a sound Leibnizian principle of identity through change for some one kind of person (human-person, dolphin-person, parrot-person or whatever). There would be no one real essence of person as such; but every person could still have the real essence of a certain kind of animal. Indirectly, this would be the real essence <u>in virtue of which</u> he was a person. (Wiggins 1980: 172)

Since Wiggins sees a natural-kind element as part of the definition of "person", he feels justified in setting fission thought-experiments aside. It is worth stressing that Wiggins's argument is aimed specifically at <u>fission</u> thought-experiments, and not at thought-experiments in general[3]. Even

if his arguments were successful, then, it would only be fission cases that would be affected, and not the method itself. Nevertheless, I wish to defend the use of fission cases, and the limited scope of Wiggins's argument has some importance for this end. Since Wiggins's argument concerns only fission cases, it is still legitimate for us to use thought-experiments not precluded by his argument. We would at least not be guilty of begging the question against him.

The crucial point in Wiggins's case against fission is his insistence that the concept of a person has a natural-kind component - that "every person would belong to some natural kind" (Wiggins 1980: 172). This is crucial since it is the laws governing the relevant natural kinds (humans, dolphins, etc) that rule out fission. It is against the claim that persons have an essential natural-kind element that (non-fission) thought-experiments provide us with a powerful argument.

We do not need to look far for the relevant imaginary cases. In Chapter 3 section 3 I suggested that cartoon characters provide evidence against Flew's model for the meaning of terms like "person", and they do damage to Wiggins's argument as well. Cartoon situations often present persons who are members of natural kinds amongst whom we never actually find persons, and to whom Wiggins's anti-splitting laws might not apply. For instance, should the cartoon character be an amoeba, then precisely what it can do is split. When the characters are artefacts of one kind or another, the point against Wiggins is even clearer: we have

examples which nobody hesitates to view as persons which are not members of any natural kind.

Putnam puts forward similar considerations using a science-fiction example:

> robots can be inspiring or pathetic - they can overawe us with their superhuman powers (and with their greater than human virtue as well, at least in the writings of some authors), or they can amuse us with their stupidities and naivete. Robots have been "known" to fall in love, go mad (power- or otherwise), annoy with oversolicitousness. At least in the literature of science fiction, then, it is possible for a robot to be "conscious"; that means (since "consciousness", like "material object" and "universal", is a philosopher's stand in for more substantial words) to have feelings, thoughts, attitudes and character traits.
>
> (Putnam 1964: 386)[4]

Robots clearly do not form a <u>natural</u> kind, but we have no problem with ascribing to them all the important features of personhood. Without some strong argument to the contrary, then, we have no reason to take seriously the natural-kind clause in Wiggins's definition of a person. And once the natural-kind clause is given up, Wiggins's argument against fission collapses.

SECTION 5: *A refinement of Wiggins's rejection, and a response*

A solution has been suggested which appears to serve Wiggins's purpose of justifying the ignoring of problematic thought experiments like fission, yet without casting person as a natural kind. Kitcher has proposed (Kitcher 1979) dropping the attempt to make person a natural kind, but at the same time looking elsewhere for laws which will determine personhood and personal identity in the way in which Wiggins envisaged natural laws working. Kitcher contends that even though no laws of natural science bring together the class of persons, there is a body of true empirical generalizations which does.

These are the generalizations of commonsense or folk psychology, which have already cropped up in chapter 2: laws (or, for the fussy, "laws") such as: "An individual who wants p and believes that doing A will bring about p, will tend to do A", and "An individual who fears q will tend to avoid situations in which that individual believes q to be likely to occur". Intentional predicates like those in these examples will be central to the theory which serves to define persons, but there will also be other laws such as "An individual who has a painful experience with objects of a certain kind will tend to eschew objects of that kind" (Kitcher 1979: 545). These are the generalizations which we use to explain the behaviour of others, and as suggested before are like natural laws in being universal, projectible, counterfactual-supporting, and so on. They do differ in being liberally

sprinkled with qualifying phrases like "tends to", and in having a large number of ceteris paribus clauses, but they are not being claimed to be the laws of a mature scientific theory, and this vagueness does not prevent them from performing their task of explaining behaviour.

There is thus no prima facie reason why folk psychology cannot perform the extension-fixing task assigned to the theories of natural science. Although the kind thus defined would not be a natural kind, it would nevertheless be a "law-governed" kind (Kitcher 1979: 544). On this model, a person would be a member of the class of individuals of whom folk psychology is true - individuals whose behaviour can be explained and predicted by the laws of folk psychology. This move also lets in certain individuals not allowed by Wiggins but which, despite his qualms, probably should be there. It is just chauvinism simply to rule out complex computers as possible persons, and perhaps corporations should be included as well (French 1983); on this model, these could be persons.

Treating person as a law-governed kind like this is in the spirit of Wiggins's defence: note the occurrence of the central terms of folk psychology in his functional specification. It also seems to solve the problems he faced in the following ways. Like the natural kinds strategy, it provides a justification for our eschewing scenarios involving the fission of persons, and so on. This strategy aims to retain the basis of the relevant concepts in the actual. Thus,

> unless actual persons divide or merge, we may ignore questions about whether clones would be persons, and if so, how many persons. (Kitcher 1979:546)

Kitcher does not explain the detail of this argument, but presumably it is analogous to Wiggins's - that such occurrences are not in accordance with the laws of folk psychology describing the normal development of an individual, and thus remain irrelevant to the inclusion and identity conditions of the kind "person".

This defence of Wiggins's strategy by appeal to law-governed rather than natural kinds may seem to solve the question of the legitimacy of "fission" thought-experiments. But there is good reason to believe that it is not at all adequate for performing this task in the way envisaged. In the remainder of this section it will be argued that folk psychology does not really preclude the envisaged scenarios as required.[5]

Does folk psychology really have laws that make fission and body-swapping impossible? Is it not rather our cognitive dependence on the laws of physics that leads us to doubt these possibilities? For folk psychology, as its detractors point out, has been around more-or-less unchanged for centuries, co-habiting happily with worldviews peppered with disembodied persons and the like. Whether you are an angel or a human, or whether you are a splitter or a non-splitter, if you believe that doing A will lead to p, and you want p, you will tend to

do A. Certainly we may worry about how a spirit could do certain kinds of A consistently with the principle of the conservation of energy; but that takes us way beyond the strictures of folk psychology. The appeal of Wiggins's natural-kind strategy was that it assigned persons a physical essence, and it was precisely that which militated against fission; by moving from the physical to the psychological as Kitcher does, this crucial element of the strategy is lost.

Perhaps this misses the central point raised earlier in connection with Wiggins's reasoning against troublesome thought-experiments. On both the law-governed and natural kind models, it will be laws governing the typical history and characteristic development of members of the kind person which make fission and so on impossible and thus irrelevant. But does folk psychology have such laws? And if it does, do they really rule out fission and body-swapping? It was argued in the previous section that Wiggins's appeal to typical human development was unhelpful in response to thought-experiments like that outlined by Parfit; here I will argue that when it comes to typical development according to folk psychology, the case against thought-experiments is even worse.

Laws governing typical development which might form part of our folk psychology would presumably concern the way a person's mental outlook matures, the different kinds of ideas and attitudes individuals tend to have at different ages, and so on. It's not at all clear, however, that these laws militate against fission: it seems rather to be the case that they could still apply despite the division of one person into

two, or despite a change of body. Even though it's true that the actual typical development of humans does not involve splitting into two, it is not this aspect of development that folk psychology describes.

An important point which has emerged in discussion of folk psychology in the literature is that it is a "topic-neutral" theory (Smart 1959; Putnam 1980; Lewis 1980): it sets out the roles mental states play without specifying the first-order properties of those states. As a result, it is consistent with both physicalist and dualist views about the nature of those states. This compatibility with dualism is of import in our discussion, because it allows for the possibility of a separation between persons and their bodies; it allows that persons can be distinct from their bodies, and thus at least prepares the ground for disembodied and dividing persons. In the light of this, it emerges that the use of the laws of folk psychology to fix the extension of the term "person" is the wrong sort of strategy to follow in attempting to avoid the thought-experiments in question. Folk psychology is just not the right theory to appeal to in trying to show the envisaged scenarios to be irrelevant and ignorable.

In these ways the law-governed kind strategy turns out to be no improvement at all on Wiggins's natural kind strategy as an argument against fission thought-experiments. Wiggins's strategy shows flaws in its attempt to set out what makes persons members of that kind, while the former strategy is also unable to perform the task of ruling out popular thought-experiments for more important reasons, as just discussed. As

a result, these attempts to show fission and other thought-experiments to be irrelevant to the debate on persons fail.

While they fail in this regard, the law-governed model makes significant progress as an account of how the semantics of "person" works, and links up neatly with the arguments of Chapter 3. In Chapter 3 I set out how the causal theory seemed at first glance to support Fodor's conclusion about the semantic irrelevance of thought-experiments: the latter reveal things about our beliefs or implicit principles, and the causal theory casts our beliefs as semantically irrelevant. I argued against this that our beliefs are not irrelevant to the semantics of "person", and here we have further support for that claim.

The law-governed model brings to the fore that the theory which does the work of fixing the extension of the kind around a paradigm person is a folk theory, the theory of folk-psychology. As a folk theory, it is one which everybody knows. This marks the case of "person" as crucially unlike the natural-kind terms for which the causal theory of reference was designed; the theories which fix the extensions of those terms are the theories of mature science, which are by no means available to everyone. It was precisely because the extension-fixing role for natural-kind terms is played by such theories that our beliefs appeared irrelevant and Fodor's conclusion seemed to get support; but once the asymmetry between the natural-kind case and our case is realized, this support vanishes. The outcome is that the method of thought-

experiment has not been shifted from its central place in the personal identity debate.

NOTES

1. Apart from the arguments I discuss in this and the following chapter, this view is also expressed by Baillie (1990), and a version of it is put forward by Johnston (1987).

2. Nagel also proposes treating person as a natural kind (Nagel 1986: 39-40). He sees this as a way of ruling out the dilemma about indeterminacy which is a consequence of fission thought-experiments. My argument in Chapter 7 addresses the indeterminacy problem, but without Nagel's manoeuvre.

3. Wiggins does take exception to some other thought-experiments besides fission ones. He writes, "those who use the method of possible worlds to determine answers to questions of necessity and identity or necessity and origin are using a mistaken method" (Wiggins 1980: 213). But the reasons he gives there are not the same as those under consideration in this chapter. I pay some attention to these new arguments of Wiggins in Chapter 8.

4. Putnam is admittedly hesitant about drawing the conclusion that we must say robots are conscious on the grounds of this sort of experiment. But in the end he admits that it is no more than chauvinism which can prevent us from making that assertion (1964: 407).

5. An interesting point can be raised here. This concerns the possibility of the radical falsehood of folk psychology. Eliminative materialism - a fairly popular theory in contemporary philosophy of mind - claims precisely this: the claims of folk psychology are false; terms like "belief" and "desire" fail to refer to anything in the world (Churchland 1981, Stich 1983). I do not wish to get embroiled in this debate here, but it is worth noting that if the eliminative materialists are right, then if we follow Kitcher's line, there is no such thing as a person. For the purposes of this thesis I will assume that the claims of folk psychology do not fail of reference. There are, in fact, solid grounds for this assumption. On this, see Horgan and Woodward (1985).

CHAPTER 5: WILKES AND PEOPLE WHO SPLIT

SECTION 1: *Wilkes's rejection and an initial reply*

In the first chapter of her book Real People Wilkes puts forward a strong case against the use of thought-experiments in our context. Her argument has a number of links with that of Wiggins, but also brings new considerations to bear on the matter.

One similarity is that the very thought-experiments which bothered Wiggins - those involving the fission or splitting of persons - are the ones which bear the brunt of Wilkes's attack. She argues (Wilkes 1988: 37) that while some philosophical thought-experiments merely need important and drastic changes to be made to them, "fission" ones must be rejected out of hand. The case she offers against all thought-experiments is dependent on the case against fission ones, and for the purposes of this discussion I will confine my attention to this central part of her argument. I wish to argue that none of the considerations she brings to bear should lead us to give up the use of this thought-experiment. In Section 1 I will outline Wilkes's case against fission and point out why, on the face of things, it is unsuccessful. In

Section 2 I will look in detail at how Wilkes hopes to avoid these prima facie problems, and I will explain why her attempts at reconciliation do not work.

Wilkes's argument runs as follows. Any thought-experiment, be it in science or philosophy, presupposes that all the relevant background conditions for the phenomenon in question are included and specified. Wilkes calls this "establishing the phenomenon". Thought-experiments are counterfactual in that they require us to suspend belief in some way: we are asked to imagine what would happen, or what we would say, if something which is not actually the case were to happen. But the background against which this change is to be supposed must be provided if the enterprise is to have any point and hope of success. In the context of personal identity, we are out to discover "the heart of our current, present notions of what it is to be a person" (1988: 12), and if we are to do this we need to know enough about the situation we are being asked to consider.

In the case of one person splitting, amoeba-like, into two, Wilkes argues that we are just unable to fill in the background in any adequate way. There are two broad ways in which the phenomenon of fission fails to be established, and they can be summed up as the difficulties of imagining, or describing, how the split occurs and of imagining that the split occurs. Wilkes argues, as did Wiggins, that the natural laws governing human beings preclude fission: it is theoretically impossible that a human divide into two. That

is the problem of describing how a person splits: against the background theory of human beings it is just not possible.

Besides this problem there are other grave difficulties in store; for, Wilkes contends, in order to imagine that a person divides or to respond to the question of what one would say if a person were to divide, one needs to know all sorts of things about, for instance, the predictability of splitting and the social background against which it would occur. These are details, however, which the proposers of the experiment do not and cannot provide.

Does all this show that we should eschew thought-experiments involving fission, and that we can safely ignore any conclusions which they seem to yield? Wilkes thinks so, but has the point really been made? Consider first the argument from the premise that it is theoretically impossible for humans to divide. It was suggested in Chapter 4 that there is reason for doubting that our current scientific theory constituted by laws governing human capacities, features and normal development, precludes a human being from splitting into two. And even so, the conclusion of the argument is that it is impossible that <u>persons</u> divide, or at least that we can ignore thought-experiments in which they are described as doing so. It would be absurd to deny that most, if not all, of the persons we actually meet are humans, or that the paradigm of a person in anyone's view would be a human, but that does not prevent there being a distinction between person and human, as is acknowledged by Wiggins, and which stops Wilkes's inference from going through.

The suspicion that this is the case becomes more pressing in the light of part of Wilkes's own argument. For she allows that there may be individuals in other galaxies who are persons, but not humans, and can split (1988: 36). Surely if this is a possibility, then arguments from the human condition in the context are irredeemably weakened? All they would show is that some persons cannot split - true, those non-splitting ones are the ones that we persons happen to be - but this consideration will not be of any help on our way to answering deep questions about the nature of personal identity.

Wilkes is always insistent on stressing that the only important impossibilities in thought-experiments are those relevant to the aims of the experiment. Thus it is that Einstein's famous experiment which set out to answer what someone travelling beside the front of a beam of light would see is not ruled inadmissible by the impossibility of a human travelling at the speed of light: this impossibility is irrelevant to the aim of the experiment. The aim of fission scenarios are, in her own words, to reveal "the heart of our current, present notions of what it is to be a person". If our current, present notions allow splitting persons - albeit alien ones - then it is not at all clear that such things are irrelevant impossibilities. The impossibility (if it is one) of human fission would rather seem to be the impossibility which is irrelevant to the outcome of experiments in which the aim is to find out about persons as such.

SECTION 2: *Why tolerating splitting does not forsake philosophy for fairy story*

This is not how Wilkes sees things. What her response would be can be seen from the points she makes after allowing the possibility of non-human divisible persons. She holds that to take notice of such purported individuals is to cross the divide between philosophy and fairy story: to indulge in an enterprise which cannot yield the results initially required.

There are two points which can be gleaned from Wilkes's discussion in support of this claim that to take an interest in dividing non-human persons is to forsake philosophy. The first involves the charge that the kind of possibility at stake here is an uninteresting one (1988: 18); the second concerns the precise aim of the thought-experiment enterprise (1988: 37).

With regard to the kind of possibility operating in the case of dividing alien persons, Wilkes allows it to be clear that the situation is logically possible, but she denies that bare logical possibility is of any interest. For it is also logically possible, i.e. consonant with the laws of logic, that water is not H_2O, and that iron bars float on water[1], and so on, but this bears no relation to what might or could happen. On the contrary, says Wilkes, we <u>know</u> that water cannot be anything but H_2O in the sense that we cannot believe in the actual existence of something which is water but not H_2O, or something which is an iron bar, but floats on water. "There is, we can say, no possible world in which these states

of affairs obtain" (1988: 18). It is thus only in the context of fiction, where we willingly suspend belief in order to be entertained, that situations such as these, and of persons splitting, have any place. The aim of the philosophical thought-experiments is to "establish" the required phenomenon so as to draw conclusions about the nature of persons, and in so far as that aim is concerned, the situation is impossible.

The second point, which is related to the first, is this: the question originally posed by the thought experimenter concerns "what we would say if we divided or fused" (1988: 37). What is to be said if aliens divide is their philosophical problem and not ours - not, at least, until we actually come across them.

I wish to argue that neither of these points pushes the thought-experiment in question over any important divide between fact and fiction. Consider the first argument in which division of persons is compared with the floating of an iron bar or water's not being H_2O. By analogy, just as it is impossible for iron to float on water, or for water not to be H_2O, so persons cannot divide. There may be no logical inconsistency here, but we know that none of these could happen. To all intents and purposes, they are impossible.

But this analogy does not work. The reason that iron will not float on water is that it has a specific gravity at 20°C of 7.87, and this precludes a bar of it from floating on water. Specific gravity is a function of the inner structure which is essential to something's being iron. Likewise, water cannot but be H_2O because that is precisely the structure of

the molecules which provides the essence of water. The other examples of impossibilities which Wilkes provides (1988: 18) also depend on essential inner structures. The problem is that the essence of person, whatever it is, is not a matter of inner structure, unlike that of iron, water, and the like. And certainly, Wilkes has no argument to the contrary. A point which emerged in the discussion of Flew's arguments as well as in the accounts of Wiggins and Kitcher is that "person" is a term inextricably linked to folk psychology, and it is from this context, at least in part, that it derives its meaning: persons are the things which have beliefs and desires, self-consciousness, and so on. If any laws apply to them and serve to constitute a kind, they are laws about the relation between beliefs and behaviour, the relations between the different kinds of pain one may feel, and such things. These laws are not in any essential way about the inner constitution of anything. Folk psychology works in complete ignorance and independence of the inner workings of the individuals to which it applies.

Wilkes herself seems to acknowledge this point when she appeals to Dennett's six criteria as the "conditions of personhood" (1988: Ch 1 sec 6; Dennett 1976) and when she acknowledges that "it is improbable that 'person' is in any legitimate sense a term that can usefully be construed as a 'natural kind'...; whereas 'human being' seems to be just that" (1988: 15). But given that the kind we are concerned with is crucially unlike the kinds Wilkes appeals to in demonstrating interesting kinds of possibility, her argument

gives us no reason to believe that the division of persons is to be ranked alongside the happenings in Alice in Wonderland as regards its informative status.

There would be a problem if the scenario or world outlined in a thought-experiment somehow conflicted, either explicitly or implicitly, with what we know to be essential features of the kind in question. If such was the case, then in Wilkes's terms the phenomenon would not have been established. But this is precisely what is not clear in the case of persons dividing, and which is clearly so in Wilkes's other examples. If the fiction being considered seemed to us to be incoherent, we would then have some evidence that it was not constructed in conformity with our ordinary concepts, but once again this is not to be observed in the case of fission. Even if the details of the division are left vague, there does not seem to be anything in folk psychology and the more familiar features of persons qua persons which suggests a conflict with division, as was argued in chapter 4, §3. Indeed, this seems to be precisely the reason why we understand fission thought-experiments so readily, and intuitively feel their importance regarding questions of identity.

The second point Wilkes makes in favour of fission being only fiction does not provide much help to her overall case. This was the point that what is at stake in the thought-experiment is what we would say if we were to divide. It does not help because this is not what is at stake. Or, at best, this is a misleading description of what is at stake. The

question is rather what we would say if persons were to divide. Now this is precisely of interest because we are persons, but to insist that we must confine our reflections to the possibility of us humans dividing is to change, if not to beg, the question. We ask the question to investigate the conceptual consequences of persons dividing. It is enough to set up the investigation that some of our fellow persons can divide, and this Wilkes acknowledges, just as she acknowledges that the concepts of human and person are different ones. That some persons do not divide because they are also humans is neither here nor there; as long as persons can divide - that is, as long as an individual can be a person and divide - then the resulting conceptual issues affect us persons too, even if we human persons don't go about splitting.

SECTION 3: *Imagining that people split*

It still remains to examine the other general thrust of Wilkes's case - that which concerns and questions the "that", as opposed to the "how" of personal fission. Wilkes argues that certain details crucially relevant to the aims of the experiment are inevitably left out of the description of the supposedly possible world. She claims that where the case of division is concerned we need to know such things as

> how often? Is it predictable? Or sometimes predictable and sometimes not, like dying? Can it be prevented? Just as obviously, the background society, against which we set the phenomenon is now mysterious. Does it have such institutions as marriage? How could that work? Or universities? It would be difficult, to say the least, if universities doubled in size every few days, or weeks, or years. Are pregnant women debarred from splitting? The <u>entire</u> background here is incomprehensible. (Wilkes 1988: 11)

Presumably the idea behind the claim that we need to know these things is that we require such information in order to be in a position to provide a sensible and informed answer to the question of what we would say if one person divided into two. That seems to be why the questions are "crucially relevant to the aims of the experiment". We would just not know what to say unless this background was filled in. Since thought-experimenters are not about to do the filling in, they cannot really expect an answer at all.

The questions Wilkes asks in the quoted passage are certainly interesting ones, which do raise fascinating puzzles with regard to splitting persons. But she consigns them far too much weight in the context in which she raises them. For consider, is it really crucial to know whether splitting is inducible or predictable in order to respond to the question "who is this person?" after a split? Likewise, would one have to know how marriage operates in the speculated person's society? It would only be necessary to know these things if

knowing what one would say to splitting requires knowing how one would live with the phenomenon. In outlining the method of thought-experiment Wilkes makes it clear that she requires this:

> Such a question typically postulates an imaginary state of affairs, something that does not in fact happen in the real world. Put another way, in the modern jargon, we imagine a "possible world" in which the state of affairs actually occurs...Then we try to draw out the implications - "what we would say if" that imagined set-up were to obtain; that is, if we inhabited that possible world.
>
> (Wilkes 1988: 2; my italics)

But in equating knowing what we would say to splitting with knowing how splitters think Wilkes seems to be guilty of the very mistake I attributed to Fodor. That is, she confuses the question of what (given our conceptual scheme) we would say to splitting with the question of what the conceptual system of a society of splitters would be like. As we saw in Chapter 3, an answer to the former question does not require an answer to the latter.

Wilkes's questions all concern how a society in which splitting occurred would cope with that peculiarity. This suggests that she is guilty of the same mistake as Fodor, but we can also use the discussion of Chapter 2 section 3 to show why her questions are misguided or irrelevant to the task of

investigating persons and their identity. It was suggested there that the concept of a person and that of personal identity get their content from their place in the system of folk psychology. Now, given the familiar principles of folk-psychology, it is not at all incomprehensible how marriage (for example) would work in a society of splitters. Perhaps the splitters would have a new kind of extended family - a family would not only grow by producing children, but by producing adults as well. Perhaps the splitters would fight for the undivided spouse. These and other possibilities are all perfectly compatible with our system for explaining behaviour.

In a similar way, questions about the frequency or predictability of splitting can be seen as irrelevant. Perhaps fission takes place once every six years but does not happen to pregnant women and reliable symptoms signal its imminence. Perhaps it happens far more often in a more haphazard way. Either way, there is no conflict with the system in which the terms "person" and "same person" have their place, and thus Wilkes's argument misses its mark. It would only be in a very peculiar thought-experiment that we would need to know that x's splitting was predictable in order to know whether x is or is not y. Whether our intuitive responses to a thought-experiment situation are dependable, whether they are really what we would say, and how we know that, are all questions which may affect the acceptability of a thought-experiment - but these are all matters independent of the knowledge that Wilkes specifies.

The points made in the previous paragraph can be reinforced by looking at another issue concerning the background society which Wilkes raises. In regard to how universities would work in the splitting society, she comments that things would be "difficult, to say the least". But that's just it - these are difficulties for the society in question, difficulties which are the consequences of (metaphysical and) other more fundamental features of the society. It is a practical difficulty with perhaps important moral implications how marriage would work in a society of splitters, but these are problems subsequent to problems of identity. It may well be difficult and even impossible to live as we do in such a society. (This, I have argued elsewhere [Beck 1989], would be the case for a society of Parfit's persons.) But that doesn't mean that persons cannot divide or that we can ignore arguments concerned with the consequences of division.

Wilkes's arguments against fission fail to establish the irrelevance of such thought-experiments to the debate on persons and their identity. Experiments outlining some scenario in which one person divides are intuitively perfectly intelligible, as well as being deeply puzzling, and as a result there is an onus on Wilkes to show why we would be justified in ignoring them. The arguments above demonstrate that she has not succeeded in this. As a result, her arguments get us no further than those of Wiggins and Kitcher, and fission remains a problem.

NOTES

1. This example, like the other ones Wilkes uses, is drawn from Seddon (1972).

PART TWO

WHAT THOUGHT-EXPERIMENTS CAN SHOW ABOUT US

CHAPTER 6: THOUGHT-EXPERIMENTS, THE SELF AND THE FUTURE

> These bizarre fictions have their uses in abstract studies, as aids to a better grasp of the nature of our ideas. (Leibniz 1765: 314)

SECTION 1: *Williams's two presentations of the experiment*

One of the most influential thought-experiments in the literature is the one which occurs in Williams's celebrated paper, "The Self and the Future". As with other thought-experiments, there is disagreement as to precisely what this particular case shows, but most commentators have found its consequences to be extremely significant for the debate on personal identity.

In this chapter I will set out Williams's thought-experiment and then discuss some of the conclusions which have been drawn from it. My hope is that something important will emerge from this central example about what thought-experiments can do and can be expected to do. This aim has two parts to it, one methodological and one substantive: my interest is in both what the experiment shows about the method

of thought-experiment and what it shows about personal identity itself.

The thought-experiment is set out as two distinct scenarios, but its force turns on its being revealed to be one and the same scenario differently described. The first description is a fairly straight-forward account of a thought-experiment of the kind used in arguments for a criterion of identity in terms of psychological continuity.

A machine has been created which is able to extract and record all of the information stored in one's brain which is relevant to and determines one's mental life. Two individuals A and B are subjected to this process, and the information from A's brain is then fed into B's brain and vice-versa. After the process, the person in B's body seems to remember having A's experiences, has A's beliefs, desires, projects, emotional attachments, and so on. Likewise, the person in A's body has the psychological features previously associated with B. The obvious intuitive response to this scenario is that a body-swap has occurred: the person in the A-body is now B, while A occupies the B-body. This intuitive response strongly supports the view that our concept of personal identity turns on psychological continuity, and that physical continuity is not necessary for identity.

Williams re-inforces this conclusion by considering the likely responses of the individuals involved. As Williams sets the case up, A and B are told about the process that is to be carried out on them and are told that one of the emerging persons will be tortured and the other rewarded.

They must each make a choice beforehand as to which is to receive which treatment, presuming that this will be done on purely self-interested grounds. The support for the psychological criterion comes from the judgements the two persons would make as to how wise their earlier choices were once the operation is complete and the torture and reward handed out. For instance, had A chosen that the A-body person be tortured and the B-body person rewarded, the B-body person would seem to remember making this choice and, if this is indeed what happens, he would be satisfied that "his" choice was the wise one (Williams 1970: 48-49).

Further re-inforcement comes from considering other choices:

> Suppose that A chooses that the A-body-person should get the money, and the B-body-person get the pain, and B chooses conversely...The experimenter announces, before the experiment, that the A-body-person will in fact get the money, and the B-body-person will get the pain. So A at this stage gets what he wants (the announced outcome matches his expressed preference). After the experiment, the distribution is carried out as announced. Both the A-body-person and the B-body- person will have to agree that what is happening is in accordance with the preference that A originally expressed. The B-body-person will naturally express this acknowledgement (since he has A's memories) by saying that this is the distribution he chose; he will recall, among other things, the experimenter announcing this outcome, his approving it as what he chose, and so forth. However, he (the B-body-person) certainly does not like what is now happening to him...The A-body-person will on the other hand recall choosing an outcome other than this one, but will reckon it good luck that the experimenter did not do what he recalls choosing. (Williams 1970: 49-50)

Once again, the psychological criterion is supported. The choices A and B made were those which one would expect from adherents of the physical criterion and, as Williams comments, "in this case the original choices of both A and B were unwise" (1970: 50).

The second description is set out in a different manner. We are asked to consider a series of cases, each a development on the previous one, and challenged to state at which step some crucially relevant change - a difference which could amount to a change of identity - occurs. The first case is one in which A is operated upon in such a way that he loses all of his memories. Williams suggests that this change will not be sufficient to support a judgement that A has lost his identity. The most important consideration here, according to Williams, is that if A were told that after the operation his body will be tortured, A would still have reason to fear that torture despite the intervening memory-loss. These are the six steps which develop on this one:

(i) A is subjected to an operation which produces total amnesia,
(ii) amnesia is produced in A, and other interferences lead to certain changes in his character,
(iii) changes in his character are produced, and at the same time certain illusory memory beliefs are produced in him; these are of a quite fictitious kind and do not fit the life of any actual person,
(iv) the same as (iii), except that both the character traits and memories are designed to be appropriate to another actual person, B,

(v) the same as in (iv) except that the result is produced by putting the information into A from the brain of B, by a method which leaves B the same as he was before,
(vi) the same happens to A as in (v) but B is not left the same, since a similar operation is conducted in the reverse direction. (Williams 1970: 55-56)

At stage (i) we have no difficulty in agreeing that A survives the loss of memory; that is, the person who emerges from the operation is identical with A. Williams suggests that there is no relevant difference between stage (i) and stage (ii) which would justify a judgement that the person emerging in (ii) is not A. Furthermore, this holds for all the following stages as well. The crux comes at the final stage. Since no line can be drawn between any of the stages, we are obliged to say that the person who emerges here is A; but the problem is that the scenario set out in (vi) is precisely that described in the first experiment. In stage (vi), just as in the first experiment, the psychological features of A are transposed into B's body and vice-versa - only there our intuitions told us that the emerging person was B.

The result is that one and the same thought-experiment merely described in different terms evokes directly conflicting intuitive responses. The one response appears to support a psychological criterion of identity, the other a physical criterion, yet these two criteria are mutually exclusive.

According to Williams's exposition, then, we are faced with a conundrum. But precisely what the effects of this

conundrum are is a question which needs careful investigation; certainly there have been many divergent responses expressed in the literature. One reaction has been to argue that the conundrum is only apparent, and that in fact there is some fault with one or the other, or both, of Williams's scenarios. In this way the puzzle would be defused, and need establish nothing radically new about identity or thought-experiments. Williams's final response, although tentative, is along these lines; he suggests that one should "take the risk" of accepting that identity goes with bodies (1970: 63).

Another response has been that Williams's thought-experiments both succeed, but that the consequent position we should adopt on personal identity is not Williams's tentative affirmation of a bodily criterion but a currently unpopular nonreductionist view.

Perhaps the most important reaction for our concerns is a rather different one which sees the consequences of the conundrum as being primarily methodological, the argument being that the stalemate which the thought-experiments bring us to shows that the method of thought-experiment is fatally flawed as a way to reaching answers about our identity. In this chapter I will take a look at all three arguments and formulate my own response to the problem, both as regards methodology and personal identity itself.

SECTION 2: *Noonan's argument that both thought-experiments are flawed*

Can we just dismiss Williams's conundrum? Noonan argues (Noonan 1989: Ch 10) that we can ignore any apparent consequences, methodological or other, of "The Self and the Future", because both the thought-experiments presented there suffer from drastic internal problems. From the way they are described we can infer that they do not support the conclusions suggested, and certainly do not lead to the conundrum that Williams sees. This argument is not one against thought-experiments in general; it is intended specifically against those outlined above.

Noonan argues against both of Williams's experiments: that the first does not support a psychological criterion of identity, nor the second a physical one. I do not wish to enter this debate at this stage, although my final response to Williams will show that I am sympathetic with one aspect of Noonan's argument. I will argue, more immediately, that Noonan's arguments against both legs are flawed, and will then go on to look at possible consequences of the success of the experiments.

As regards the apparent body-swap outlined earlier, Noonan finds that it fails to support a psychological criterion because of the way in which Williams plays down the crucial role of certain of the participants' (A's and B's) psychological attitudes.

The problem becomes clear, according to Noonan, when one looks at the various stages Williams follows in support of his

conclusion. Williams considers A's and B's reflections on the wisdom of their earlier choices made according to (i) a psychological criterion (Williams 1970: 48-49), (ii) a physical criterion (49-50), and (iii) different criteria (50) - where A chooses according to a psychological, and B a physical, criterion. In each case, as outlined above, Williams finds that the individuals' reflections would support psychological continuity as the criterion.

Noonan argues that what Williams must do in order to conclude thus is to ignore, or at least drastically play down, his own assumptions. For how the individuals rate earlier choices will depend crucially on their beliefs as to identity criteria. Take case (ii) for instance. If A and B choose as they do because they accept a physical criterion (i.e. A that the B-body be tortured, B choosing that fate for the A-body) and the B-body person is subsequently tortured, it is not at all clear that this person (B-body) will acknowledge that his original choice, the one he seems to remember making, was mistaken. Rather, suggests Noonan, he will dismiss this apparent memory as illusory - for both A and B believe in the physical criterion - and will complain that the torture was not meted out as he, <u>that is B</u>, chose. Williams, in suggesting what he does, ignores the force that the fundamental beliefs which shape the participants' choices will have when they come to reflect on those choices. Since he does this consistently throughout the discussion of the first experiment, there is no support to be gleaned there for the psychological criterion.

It is because of this supposed misrepresentation of how those involved would react that Noonan takes Williams's first description of the thought-experiment to fail. However, it is not at all clear that it is Williams and not Noonan who is guilty of misrepresentation. Williams's suggestions as to how A and B will react after the experiment are based on the intuition that individuals will take themselves to be the persons they feel like, and whose lives they remember. This seems to me to be fairly uncontroversial, even when it comes to an individual who believes strongly in the physical criterion of identity. For it certainly does not follow from the fact that A overtly adopts the physical criterion that if A were to look in the mirror and see B's body, he would say, "Oh look, I'm not the person I think I am!" Indeed, it is most implausible that A would react in this way, yet that is precisely what Noonan is suggesting. As a result, Noonan's case against the first experiment is anything but convincing.

Noonan also argues that the second experiment fails to establish its conclusion. This is because he claims that a line <u>can</u> be drawn between two of the steps Williams outlines, namely between (iv) and (v). At this point, according to Noonan, one who holds that psychological continuity is necessary for identity can insist that the person who emerges in (v) need not be identical to that in (iv) without contradicting intuitively correct claims about survival through amnesia, and so on.

The difference is that in (v) the apparent memories of the A-body person have been brought about by a "very special

causal process which ensures that the brain of A has been wiped clean of all the information it contained and that the A-body person is psychologically continuous with B" (Noonan 1989: 225), while in (iv) all that occurs is that illusory memories modelled on B's are induced in A. The difference is not just, as Williams claims, that in (v) the model for the A-body's apparent memories is also their cause. The crux of Noonan's point is that the process whereby A's psychology is altered is in no relevant way different from removing A's brain and replacing it, as in a thought-experiment like Parfit's "My Division" (Parfit 1984: 254-255), with one hemisphere of B's brain; in that case there seems to be a clear difference between what occurs and inducing illusory memories.

Williams objects to drawing the line between (iv) and (v) on the grounds that the A-body person emerging in (v) must be A, since the existence of an undisputed B prevents his being B. Noonan suggests that this is tendentious because the B-body persons's identity with B <u>can</u> be disputed by pointing out that what happens to B is better described as fission than as the production of a copy without claim to identity. The only claim that the B-body has over the A-body to B's identity is the continued occupation of B's body, in particular, continued possession of B's brain; but to use this alone is to beg the question.

Noonan's argument is a strong one against Williams's description of the second thought-experiment. Nevertheless, it is not enough to show that we do not face a conundrum.

This is because the example can be altered so as to avoid Noonan's criticism, and yet still present what is at heart the same argument. The way to change the case and achieve this is to increase the number of steps in the description. Instead of the six cases which Williams outlines, one can require consideration of a spectrum of cases like that described by Parfit (1984: 231-233) in which each stage represents a change of a very small degree over the stage preceding it, the changes occurring a few cells at a time. Thus we would be faced by a spectrum consisting of an indefinitely large number of cases. In each case A will be operated upon to produce some psychological change, the amount of change will increase gradually as we move along the spectrum, but the difference between two adjacent cases would be almost imperceptible. In this way, the final case in the spectrum can be that described in Williams's stage (vi), but nowhere along the spectrum does it make sense to draw a line between two stages. After all, how could a change so small represent a change in something so momentous as identity? The resulting thought-experiment presents the same sort of case which Williams intended with his second thought-experiment, but which is proof against Noonan's attack.

The sorts of consideration which Noonan presents do not show any fatally damaging internal inconsistencies in the thought-experiments under discussion. Given that Williams's case still stands to be informative, the question remains as to what it does show, and especially whether it has any methodological significance.

SECTION 3: *Does the conundrum support the unanalysability of identity?*

One influential line of argument is that which takes the thought-experiments to succeed, but to succeed in showing something other than the tentative conclusions Williams draws from them. Indeed, the strategy is taken as showing that the most popular contemporary views on personal identity are quite mistaken. The argument occurs in discussions by Swinburne (1984) and Madell (1981), and runs along the following lines.

Madell and Swinburne deny that personal identity can be analysed into some more familiar or better understood relation like physical or psychological continuity, and they argue that this view gains strong support from Williams's conundrum. Personal identity to Madell and Swinburne is a basic or simple relation which resists any breaking down into something more fundamental. To be the same person is to be the same person, and there is no more to be said. One is not the same person as one was <u>because of</u> the holding of some relation other than identity. One immediate implication of this view is that it would be possible for A and B to be distinct persons even though there is no difference between them whatsoever - no difference, that is, other than their being (unanalysably) non-identical. In Madell's terms, "I might not have existed, but someone having <u>exactly</u> the life that I have had might have existed instead" (Madell 1981: 79)[1]. On the other hand, you could have had a life totally different to the one you have led, and yet still be <u>you</u>.

These entailed possibilities are important for Madell and Swinburne's arguments in the context of "The Self and The Future" because they show the acceptance by those who adopt a nonreductionist view like theirs of the possibility of "bare identity" - of identities or non-identities which do not hold in virtue of anything. Madell and Swinburne suggest that it is precisely an unwillingness to accept the possibility of bare identity which leads Williams and others who believe personal identity to be analysable into trouble with the examples at hand. If one throws out this prejudice, and accepts that physical and psychological continuity may be no more than evidence for identity, the threat posed by the two examples is removed[2].

Admittedly the criteria we usually use to make identity judgements let us down in Williams's cases by yielding conflicting results, but that is just one of the epistemological problems we must be prepared to face; our criteria let us down elsewhere as well. This failure of normal criteria does not mean that there is any metaphysical problem about identity at stake in "The Self and the Future". In the unlikely event that the operation should occur, then one of the emerging people would be A and one would be B; we may just not know which is which.

Madell places even greater stress than Swinburne on these particular examples:

> anyone who follows the two stories that Williams tells in 'The Self and the Future' must also reject the view that one or other story is incoherent. The outstanding fact about these stories is that both of them are so compelling. We are led to understand just how it is possible for there to be these two different possibilities (i.e. distinct individuals in different worlds), and for there to be no observable difference between them. Far from this being 'utterly mysterious', as Williams claims, it is precisely the conclusion that our whole argument demands. (Madell 1981: 99)

It might make Madell's claim clearer to put it in terms closer to the examples under discussion than does this quoted passage. What the first version shows is how it is possible for an individual to retain her identity despite a total physical change. The second version shows that identity can be retained despite a total psychological change. That the two are the same situation differently described is neither here nor there - together they show how identity can be unaffected by a total psychological change and a total physical change. But this is incompatible with the view that identity can be analysed in terms of some sort of observable continuity, and is in line with Madell's prediction that any attempt at such analysis will fail. As a result, a "simple" view of identity is shown to be correct. Personal identity is a simple, unanalysable relation, something over and above mental and physical continuities.

Before one rushes to accept the conclusions Madell and Swinburne draw from "The Self and the Future", some points must be borne in mind. In the first place, one needs to be

careful about accusing one who accepts that personal identity is further analysable of confusing evidence for identity with the relation itself. For it is a perfectly rational methodological tenet which says that one does not postulate entities and relations apart from or underlying those which one can observe until there is overwhelming reason to do so. Unless there is some enormous theoretical advantage to be gained, it is best to stick with the evidence. So there need be no simple confusion going on here.

Related to this is a problem which arises with regard to Madell's point about the way in which both of Williams's stories are compelling. He certainly seems to be correct that they are compelling, and since they compel us in different directions this is something which must be explained if one is to deny that the simple view is the only acceptable conclusion to be drawn[3]. But if he is right about the simple view of identity being the correct one, then his own point also becomes puzzling. It is puzzling because on that condition, neither argument should be compelling; least of all to one like Madell who believes strongly in the simple view. For if the simple view is true, then in neither of Williams's examples are we provided with evidence necessary or sufficient for a judgement about identity; indeed there is no such thing. On their view we could have no grounds for deciding either way about the identity of the persons who emerge from the operations, and so any compulsion either of the experiments carries appears to be totally incomprehensible.

While this is a problem for Madell's case, I believe there is a way of avoiding it. The challenge is not completely fair because nonreductionists like Madell and Swinburne allow that factors like psychological or physical continuity count as <u>evidence</u> for identity. It is just that those factors will not always be decisive; they may be present even in the absence of identity, and absent in its presence. They may mislead us in certain cases, as they mislead Williams into believing he faces a genuine conundrum about identity, but under normal circumstances they are good enough for us to get on with.

Neither of the two points raised in response to Madell and Swinburne succeed in undermining their case. As a result, their interpretation of the consequences of Williams's conundrum stands, at least conditionally. The condition is this: that Williams's two scenarios should be as intuitively compelling as Madell suggests they are. If they are indeed both successful and contain no internal flaws, then they do provide a strong case for the existence of bare possibility with regard to identity. In Sections 5 and 6, the question of whether this condition is satisfied will be raised.

SECTION 4: *Does the conundrum show thought-experiment to be a misguided method?*

The potentially most devastating interpretation of the two cases and their consequences is still to be considered. This would take Williams's conundrum to show that the entire method of using thought-experiments to support theories of personal identity is mistaken. If Williams's arguments go through, then what emerges is that we can have directly conflicting intuitions regarding the application of a concept in a given situation. Such a conflict might be taken as indicating that our commonsense conception of what it is to be the same person over time is incoherent. Alternatively, one might respond by rejecting appeals to intuitions about counterfactual situations, in this context at least. In that case the conflict which has become apparent is taken to show that our intuitions are unreliable as support for a theory of personal identity.

Whatever option is taken, the consequences are the same: we would have to stop using thought-experiments to support or reject alternative theories in this area. This is the position of White (White 1989). Noonan suggests that this is also the conclusion which Johnston draws from "The Self and the Future" (Johnston 1987), but I will point out later that Johnston's response is far less radical. White's response is an extreme one, with repercussions extending far beyond the area of philosophy concerned with personal identity. That may not be a reason for rejecting a response like White's to Williams's conundrum, but there are such reasons to be found.

A first important point is that the rejection of thought-experiments is only a consequence of Williams's conundrum if a very strong assumption is made: that what thought- experiments attempt to do is to elicit from us philosophically correct intuitions. For it is only if the intuitive response to each experiment is meant to be the correct response that the conflicting responses produce a contradiction. It is true that this assumption is made in much of the discussion in the literature which makes use of thought-experiments. When (for example) Locke presents us with the case of the prince and the cobbler, he believes that the intuition which "everyone sees" (that the prince swaps bodies) is the correct one. It is thus that he takes the experiment to show that sameness of body is not a necessary condition for identity. But I do not believe that the assumption is crucial to thought-experiments' providing a useful and important method of argument.

There is then a methodological moral that I wish to draw from "The Self and the Future": that we must cease to view thought-experiments as revealing the correct intuition about personal identity and its necessary and sufficient conditions. It still remains to be shown how they can serve any useful purpose once this view is abandoned.

Let us take another look at the first experiment (or a purged version thereof) and note closely what goes on there, especially the cognitive responses it invokes in us. Another thought-experiment, used in a different context, may be a helpful guide here. The experiment is one of Putnam's and

concerns, among other things, the meaning of "cat". Putnam sets out the following scenario:

> Suppose...that there never have been cats, i.e. non-fake cats. Suppose evolution has produced many things that come close to the cat but that it never actually produced the cat, and that the cat as we know it is and always was an artifact. Every movement of a cat, every twitch of a muscle, every meow, every flicker of an eyelid is thought out by a man in a control center on Mars and is then executed by the cat's body as the result of signals that emanate not from the cat's "brain" but from a highly minituarized radio receiver located, let us say, in the cat's pineal gland. (Putnam 1963: 53-4)

He then asks whether, should this be discovered to be the case, there are or ever were any cats. Does one respond that there never were any cats since being an animal is essential to being a cat, or does one respond that there are cats - and thus that cats do not have to be animals. His response is that there are and were cats, despite it now being the case that cats are not animals. However, while Putnam is clear about his own answer which is what seems to be the general response, he suggests that things are not quite as clear cut as this makes them seem (1963: 54). Unger brings out more clearly the complexity of our cognitive response in his discussion of the example:

> Notice that we do not make just one response to the question asked. Even while our dominant response is to believe that the correct answer is "Yes," we make a dominated response, of believing oppositely, that the correct answer is "No." Or, at the very least, we have a felt tendency to believe in that negative direction. (Unger 1982: 119)

Here Unger seems to be correct. We respond that these familiar objects are cats, and yet we also want to say that cats are animals; in that way, we experience the "felt tendency" against our dominant response of which Unger speaks. In the same way we have a felt tendency to deny our dominant response to talk of "demon possession" by the witch-doctor of old. Our dominant response is to say that "demon possession" occurs when a demon enters a human body, but since that never happens there is no such thing as demon possession. Our tendency to deny this (i.e. our "dominated response") shows in our wish to say something like, "demon possession is just epilepsy", which implies that there is such a thing.

In the first of Williams's thought-experiments something very similar can be observed. Our dominant response to the question, "who emerges in the A-body after the operation?" is "B." But, as in the case of Putnam's experiment and the case of demon possession, we have a dominated response to the question which is "A." In the case of the supposed body-swap the first response is far stronger, that is, our attitude is less equivocal than towards Putnam's experiment, but it is no less misleading to talk as if we have one single (and supposedly correct) intuitive response.

Unger explains responses to Putnam's case in terms of certain of our beliefs and their relative strengths: our "existence belief" that there are cats is stronger than our "property belief" that cats are animals. In the demon case, we might say that our property belief is stronger than our existence belief. A similar sort of explanation may be appropriate in the personal identity case, this time in terms of the relative strength of our belief that psychological continuity is crucial to our identity over the belief that bodily continuity is. However, talk of this sort of belief is a bit uncomfortable in the context of commonsense responses, for it is far from obvious that people other than philosophers have such beliefs.

Perhaps the best way of avoiding this worry is to describe the position of the kind of thought-experiment at stake as follows. In mastering a language, we adopt certain classificatory habits - we become disposed to classify certain individuals as "cats", "persons", "the same person as ...", and so on. These habits rest on implicit criteria according to which the classification is done; what the thought-experiments do is to make these implicit criteria explicit. They serve to bring out the principles at work, even if these principles cannot properly be viewed as fully-fledged beliefs. In some cases, like Williams's, implicit criteria which are usually co-satisfied can be made to come apart, with the possible result that we discover to which underlying principles we are more strongly attached.

To describe the results of thought-experiments as the discovery of the relative strength of commitment to underlying principles which they reveal serves to avoid unfounded faith in the existence of "the correct intuition". But it does not mean that thought-experiments have all their promise removed and become pointless. After all, these principles are constitutive of our conceptual system and we can still devise and use experiments which will decide which of two crucial and conflicting criteria or principles is the more fundamental, in virtue of being the one we are least prepared to give up or deny. Williams's example, then, can be used to show that the relation of psychological continuity is more fundamental in our conceptual system than its counterpart concerning bodily continuity. Of course, Williams's other experiment still then poses a problem, even if the problem is now slightly changed.

Even if we no longer assume that the thought-experiment will lead us directly to metaphysical truth - that is, to one clear, coherent and correct intuition, or to some implicit principle which is a strict and universal rule, "The Self and The Future" still poses a problem. For if, as suggested, the first thought-experiment reveals that our adherence to psychological features is more fundamental than to bodily ones, the second experiment contradicts this by apparently reversing the direction of dominance. As I remarked earlier, the fact that both experiments seem so compelling needs to be explained.

This problem is important, not only for Williams's thought-experiments, but for thought-experiments in general.

Should we find that thought-experiments reveal some minor conflict or tension in the implicit principles which underlie our concepts, this would not necessarily be of any fatal consequence to the method. But the existence of a case which shows a generally felt, direct inversion of intuitions like that under discussion would be a serious threat to the usefulness and reliability of the method, since it raises the suspicion that should any given thought-experiment be differently described we might feel an intuition totally opposed to the one we currently feel. In the next few sections I will be concerned to avert this threat to the method in general by averting the threat Williams's examples pose to the coherence of the concept of personal identity in particular.

SECTION 5: *What Williams's experiments <u>do</u> show*

A degree of relativity has already been introduced by giving up the search for the correct intuition, and at the same time it should be acknowledged that our responses to a thought-experiment may be affected by the way and the context in which an imagined situation is described or presented. For instance, Parfit's description of the outcome of his thought-experiment involving the teletransporter[4] makes it much easier for us to agree that we would survive the experience: he

describes it as "I" who enters the teletransporter, and "I" who wakes up on Mars in the body made of new matter (Parfit 1984: 199)[5]. And should Putnam's experiment be presented in the context of a zoology seminar, we may well react differently to Putnam. But these differences in reaction will usually not be radical ones: they may affect what we are disposed to say, yet it is extremely unlikely that they will reverse responses generally.

Williams's two experiments seem to be a special case. Even so, they do not justify the radical response outlined at the start of this section - that they be taken to undermine the method itself. They do not do so as long as some feature, or features, of their presentation can be isolated, the presence of which explains the response evoked, and as long as isolation and explanation produces a change of response. These claims are not very different from those made by Mark Johnston, where he does not reject thought-experiment as a method but suggests that our intuitive reactions to them are "defeasible judgements" (Johnston 1987: 64 and 80-83)[6].

In the case of the two experiments in question, such features can be seen in the second experiment. Perhaps the primary reason for singling out the second one as requiring explanation is that it doesn't take the straightforward form of the first. It is not as if we are simply presented with two scenarios, to the first of which we respond "A emerges in the B-body" and to the other, "A emerges in the A-body". In that way although the two thought-experiments do at some stage involve descriptions of one and the same situation, they are

nevertheless very different thought-experiments. It is not as if the second experiment is simply the first differently described. Ever since Locke's example we have found scenarios of apparent body-swaps compelling, and once the second thought-experiment is closely scrutinized, it becomes clear that we have no counter-example to that here.

How we are led to deny that A emerges in the B-body can be seen from the way in which the second case is presented. In the first place, Williams describes things as if they merely happen within the life of <u>one</u> person - amnesia is produced <u>in A</u>, illusory memories are induced <u>in A</u> - implying that it is still A who emerges after the change; but this is precisely what we are asked to judge, and our responses, I suggest, are being prejudiced by the description given.

A second misleading point is the way that the huge jump between steps (ii) and (iii) is disguised in the description. In step (ii) we are told that amnesia and "certain character changes" are produced in A. In step (iii) there are "changes in his character" and "certain illusory memories are induced". This does not sound like a very important development, certainly not enough to herald a change of identity - but it only comes across like that because the whole story is not being made clear. In step (iv) we learn that the psychological changes wrought are such that the character traits and memories shown and experienced match up to those of some other actual person, but that what goes on is essentially "<u>the same as (iii)</u>" (my italics). This means that the character and memory changes occurring in (iii) are

sufficiently different from A's to be those of a totally different, even if non-actual, person. But then what happens in (iii) is nothing like the few minor changes that Williams's description suggests, but a massive - a total - invasion of A's psychology. Now, we may have no objection to the idea of a person surviving memory loss and some character change, but a total and irreversible change like this is another matter altogether. Were the extent of the change in (iii) made clear in the description, then it is much less likely that anyone would find the move from (ii) to (iii) intuitively acceptable.

These two points are enough to show why the general response to the second experiment has been as envisaged by Williams, and are sufficient to show that that response would not be straightforwardly elicited by a clearer description of the changes that are supposedly occurring over the various stages, or at least to show that that should remain a dominated response. But the modification I suggested in Section 2 as a way of getting around Noonan's criticism of the experiment - that is, by introducing a spectrum of cases in which each case represents only a very slight change from the previous one - avoids this sort of objection. The objections just expressed are only to the way in which Williams describes the problem; a similar problem can be described in less misleading terms.

Even though the second thought-experiment can be set up in a way which avoids the problem of too large a change between its stages, I wish to argue that there remain reasons for not taking it and its apparent consequences too seriously.

The original problem is set out as a series of cases, each one representing an apparently trivial development on the previous step. This step-by-step development along with the claim that to draw a line between any two adjacent steps would be irrational is a familiar pattern of argument. The amendment proposed in Section 2 not only avoids some objections, but offers a useful insight when it comes to the way the thought-experiment shapes our responses by drawing attention to this familiar pattern operating here. For once it is recognised that we are dealing with something of the same form as a Sorites problem, we should no longer be surprised that we are led into the invidious position that step (vi) represents, nor should we see the reformulated problem as presenting any fatal threat.

With the standard "paradox of the heap", one is led by a series of small steps to a strongly counter-intuitive judgement. But one does not just accept this judgement, nor does one accept that this shows a hopeless incoherence in one's belief-system. Likewise with Williams's example, one should realise that the conclusion we are led to is <u>strongly counter-intuitive</u> in that it conflicts with some of our fundamental principles; that is what the first experiment shows, and what we have realized ever since Locke. But then, as with the case of the concept of a heap, we can accept that the most the second experiment shows is that there is some vagueness in our concept of personal identity. That is hardly a startling admission and has nothing like the consequences of the devastating incoherence with which we seemed to be

threatened[7]. I will return to the question of the possible indeterminacy of identity in some detail in Chapter 8.

In discussing some of the dangers involved in the use of thought-experiments, Unger issues the following warning which is of import here:

> Even in favourable, revealing contexts, not all examples will elicit responses indicative of philosophically interesting attitudes in the area at which the case may be aimed. Dominant responses to these potentially dangerous examples may indicate, instead, certain rather general psychological tendencies that we have. Although sometimes useful to us in other ways, for purposes of philosophical inquiry these tendencies will be distorting.
>
> (Unger 1990: 12)

In the case of the revised version of Williams's second thought-experiment, and indeed of the original version, it is, I suggest, precisely one of these tendencies that is to be detected. Faced by the use of a Sorites-like development of the argument by small stages, we tend to ignore the larger picture. If we did not have this tendency we would never react in the sympathetic way we do to Williams's second experiment, but would see it for the body-swap that it is.

SECTION 6: *Williams's own response and why it is unconvincing*

This account of how the description of the second experiment misleads us solves the problem posed by his conundrum. However, there is another argument which must be dealt with before we can move on to other thought-experiments. This argument is the final response which Williams makes to his own experiments. As I have done, he argues that one side of the experiment misleads by the way in which it is described; but unlike my case, he holds that it is the <u>first</u> description which is at fault. This is Williams's argument:

> The apparently decisive arguments of the first presentation, which suggested that A should identify himself with the B-body-person, turned on the extreme neatness of the situation in satisfying, if any could, the description of 'changing bodies'. But this neatness is basically artificial; it is the product of the will of the experimenter to produce a situation which would naturally elicit, with minimum hesitation, that description. By the sorts of methods he employed, he could easily have left off earlier or gone on further. He could have stopped at (the point equivalent to) situation (v), leaving B as he was; or he could have gone on and produced two persons each with A-like character and memories...If he had done either of those, we should have been in yet greater difficulty about what to say; he just chose to make it as easy as possible for us to find something to say...The experimenter has...produced the one situation out of a range of equally possible situations which we should be most disposed to call a change of bodies. As against this, the principle that one's fears can extend to future pain whatever psychological changes precede it seems positively straightforward. Perhaps, indeed, it is not; but we need to be shown what is wrong with it. Until we are shown what is wrong with it, we should perhaps decide that if we were the person A then, if we were to decide selfishly,

> we should pass the pain to the B-body person. (Williams 1970: 62-3, my parenthesis)

Does this leave us back in the position suggested by Noonan, able to ignore "The Self and the Future" because both descriptions are flawed? I don't think that it does, because Williams's case here is not convincing. It is not the tentativeness of his proposal which fails to convince, but rather the fact that the considerations he raises are weak ones, especially in the light of the strong case presented against the second experiment in Section 5.

In the first place, we <u>can</u> show what is wrong with the principle that one's fears can extend to future pain despite radical intervening psychological changes. The principle is only plausible if the situation is described in a question-begging way (i.e. that <u>A</u> will be tortured), and the description of the degree of psychological change is fudged, as the discussion of Section 5 makes clear.

In the second place, it is by no means clear that the will of the experimenter to produce an easy response is as guilty as Williams suggests. Certainly, the experimenter could have stopped his description at (v), leaving two people with the character and memory of B, or he could have gone on to make two people like A as well. These situations raise interesting issues, but I will argue that they do not stop one from reacting to the original experiment as one originally did. In Chapter 7 we will pay close attention to thought-

experiments that raise exactly these issues, and what will emerge there is that those experiments have no consequences which conflict with the view that personal identity can be analysed in terms of psychological continuity. Williams admits that his final answer is a risky one (1970: 63), but I believe he plumps for the wrong option.

SECTION 7: *Conclusion*

In terms of the personal identity debate, the thought-experiments described in "The Self and the Future" do not provide sufficient grounds for embracing the view that personal identity is unanalysable. Nor do they, on a more careful reading, support a bodily criterion of personal identity. Both of these positions are let down by the faults of the second experiment. Nevertheless, we are not justified (as Noonan has suggested we are) in ignoring the experiments altogether. This means that the thought-experiments ultimately suggest support for the view that if personal identity is to be analysed, it is to be analysed in terms of psychological continuity.

As far as our methodological interests are concerned, I conclude that one aspect of the popular response to Williams's thought-experiments has been correct in that they <u>do</u> have something to teach us. It is not that the method of

thought-experiment must be rejected. Nevertheless, the lesson is that we must temper our expectations of thought-experiments like those presented; they can indeed be informative, but they inform by revealing which principles underlying the application of a concept are most important to us given our conceptual scheme, not by some direct route to metaphysical reality.

NOTES

1. As mentioned in Chapter 1, Madell uses a thought-experiment along these lines as an argument for his view on personal identity.

2. In the following two chapters, independent reasons will be given for accepting this principle. However, such acceptance does not commit one to the views of Madell and Swinburne on analysability. As will be argued, there is good reason to question those views.

3. In what follows this discussion, a few possible explanations will be put forward, and there are others besides - like that suggested by Nozick's "closest continuer" theory (Nozick 1981).

4. Cited on p10 above.

5. Another fairly clear example of this sort of description influencing response can be seen in Chisholm's description and response to a case not unlike Williams's second experiment (Chisholm 1969: 104-5).

6. There is experimental evidence supporting my view that our intuitive responses are defeasible judgements, which is particularly relevant to Williams's experiments and the terms in which they are described. Kahneman and Tversky set out case studies which show how people react in contradictory ways to choices they are offered under conditions of uncertainty, the subjects' reactions depending on whether the consequences of the choices are described in terms of gains or losses (Kahneman and Tversky 1984).

7. Even the so-called paradoxes which beset vague concepts may well not be insurmountable problems. Mark Sainsbury reviews a number of possible responses in his Paradoxes (Sainsbury 1989).

CHAPTER 7 : PARFIT'S DIVISION

SECTION 1: *Parfit's argument*

A second thought-experiment which has been extremely influential and which has received fairly wide attention is one that came up in Chapters 4 and 5. This is the experiment in which one person apparently splits into two. As was mentioned, there is more than one version of this case, but I will concentrate on a single treatment: that of Parfit in his Reasons and Persons, which is the most graphic and extended version in the literature (Parfit 1984: 253-265). The thought-experiment has been claimed to have extremely important consequences for the debate about personal identity. As was the case with Williams's "The Self and the Future", our interest will be both in what these consequences are, and in gleaning what implications we can for the method of thought-experiment in general.

Before looking at the consequences of the thought-experiment, it would be useful to rehearse Parfit's account of the experiment. Parfit assumes that he is one of a set of identical triplets, and that each of his brain hemispheres is

capable of the functions of the other. He then outlines the following scenario:

> My Division. My body is fatally injured as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me. And he has a body that is very like mine. (Parfit 1984: 254)

Parfit points out that there are only four possible answers to the question, "What happens to me?" in the light of this scenario:

> (i) Nobody who survives is identical with me, that is, in effect, I have died.
> (ii) I am one of the surviving people.
> (iii) I am the other survivor.
> (iv) I am both surviving people.

The trouble is that all of these answers are problematic. It would be strongly counter-intuitive to say that the original person was dead; that is, that nobody is identical to the original. As Parfit puts it, this would amount to calling a

double success a failure (1984: 256). The logic of identity prevents both candidates from being the original. But to say that one of the candidates retains the identity of the original while the other is merely exactly similar is to make an arbitrary, ungrounded decision. However, those are all the available options.

What all this provides, argues Parfit, is the basis for a strong case in favour of what he calls a reductionist view of personal identity. The case centres around Parfit's contention that the reductionist view can accommodate the consequences of the imagined situation, while the opposing nonreductionist views can not. It will benefit the discussion which follows to look in some detail at why he sees "My Division" supporting reductionism.

Reductionism, according to Parfit, holds not only

> that the fact of a person's identity over time just consists in the holding of certain more particular facts

but also

> that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an <u>impersonal</u> way.
> (Parfit 1984: 210)

Reductionists differ among themselves as to what the particular facts in question are - whether physical or psychological or both, but those differences are not the concern of Parfit's "My Division" argument. He takes his argument to support the family of views which agree that (i) personal identity can be analysed into relations of physical and/or psychological continuity, and that (ii) these relations can be completely described without making any reference to persons or personal identity.

It is worth stressing the second part of what Parfit says here, as he himself does, for that is what makes the position described truly reductionist. Following on from claim (ii), according to Parfit a reductionist also makes the third claim that

> we could give a complete description of reality without claiming that persons exist (1984: 212).

This is a feature of Parfit's reductionism which will receive special attention in this chapter.

A nonreductionist, in Parfit's terms, is someone opposed to the broad view just described, someone who denies either both claims (i) and (ii) or who at least denies claim (ii) (Parfit 1984: 210). For the nonreductionist, personal identity is a relation which does not just consist in the

holding of physical or psychological facts. Or, at the least, a nonreductionist will hold that these facts somehow presuppose identity, and thus cannot be "impersonally described". The nonreductionist claims that it is possible for there to be a person who has a life (and body) exactly like yours yet who is not you, and conversely that you could have had a life (and body) completely different to those you do have. Nonreductionists suggest that these possibilities have our intuitive support, yet are incompatible with any attempt to reduce personal identity to physical or psychological continuity.

Nonreductionists also appeal to other intuitions in support of their position. For instance, support is claimed to come from our first-person perspective in the following way (Madell 1981:13-14). Our own identity seems to be crucially unlike the identity over time of other kinds of things. Whether or not we call a stamp collection the same collection despite the additions, subtractions and re-arrangements that occur over a period of time is very much a matter of convention. The same goes for our decision to call the old oak the same entity as the sapling and as the acorn. This is what Butler was pointing to when he described our practices as involving a "loose and popular" rather than the "strict and philosophical" sense of identity (Butler 1736: 101). But what goes for the oak does not seem to go for our own identity. It seems impossible seriously to view your own conscious life or a portion of it as a succession of discrete moments of consciousness which are only <u>by convention</u> those of one person

(viz. you). In this way, your identity is not just a matter of physical or psychological continuity, and any attempt to reduce it to that would appear to stem from some confusion. I make reference to these intuitions simply to make Parfit's target more readily understandable. In what follows, I will confine my attention to his definition, and the associated principles which he explicitly outlines and uses.

Parfit's argument for reductionism based on "My Division" takes the form of an attack on two principles which he associates with nonreductionism. He calls these the principles (i) that identity must be determinate and (ii) that identity is what matters in survival. These principles and nonreductionism "stand or fall together," he says (1984: 216), and argues that they fall. Since reductionism is compatible with the falsity of both principles, he concludes that the fall of nonreductionism is at the same time support for his reductionist view of personal identity.

The argument concerning the principle that identity must be determinate runs as follows. What emerged from "My Division" is that none of the possible answers to the question, "What happens to me?" is better than any of the others. The question is, says Parfit, an open one. What this means is that questions of identity can have indeterminate answers - the scenario of "My Division" is precisely a context in which that is the case - and any theory which implies that they cannot must be wrong. That forms the crux of Parfit's case: he holds that the nonreductionist is bound to insist

that identity must be determinate, whereas reductionism is compatible with indeterminacy.

The reductionist sees personal identity as involving nothing more than our brains, bodies and sequences of psychological or physical events. As an example, for Parfit person x at time t is identical with person y at time t-1 if, and only if, x (and nobody else) is psychologically continuous with y. Psychological connections and continuity are matters of degree, as are the physical connections to which other reductionists appeal, and thus the reductionist is prepared for indeterminacy. Questions of identity to which there are no clear answers or to which one answer is as good as another conflicting one are to be expected. On the other hand, to the nonreductionist these matters of degree are not what identity is all about - it is something over and above physical or psychological continuity. As a result, Parfit suggests, nonreductionism has no room for indeterminacy where identity is concerned. And because nonreductionism implies determinacy while reductionism does not, "My Division" is claimed to show the latter to be true and the former false.

As mentioned, there is another aspect to Parfit's argument on top of this one about the determinacy of identity, one which concerns the principle that identity is what matters in survival. Parfit explains the principle as follows.

> Consider an ordinary case where...there are two possible outcomes. In one of the outcomes, I am about to die. In the other outcome I shall live for another forty years. If these forty years would be worth living, the second outcome would be better for me. And the difference between these outcomes would be judged to be important on most theories about rationality, and most moral theories. It would have rational and moral significance whether I am about to die, or shall live for another forty years. What is judged to be important here is whether, during these forty years, there will be someone living who will be me. On one view, this is always what is important. I call this the view that personal identity is what matters. (1984: 215)

Parfit argues that what emerges from "My Division" is that the relation of identity cannot be what really matters in such a case, or indeed any other. The relationship between the original person and each of the survivors contains all that we would wish for in normal survival - it is (at least just about) as good as ordinary survival. But since the logic of identity prevents us from calling the relationship between these individuals "identity", it cannot be the case that identity is what matters most in survival.

Why this is of relevance to Parfit's case for reductionism is that, following his reasoning[1], nonreductionism implies that identity is what matters above all else in survival. Parfit assumes rather than spells out why this should be the case, but presumably his reasoning goes back to the sort of claims that were suggested above as providing intuitive support for nonreductionism and which Madell and Swinburne felt were backed up by Williams's two

cases in the preceding chapter. The suggestion was that it is intuitively plausible that you could survive without either physical or psychological continuity. If so, then it seems that neither of these can be what ultimately matter: what matters is identity. Now, Parfit's point is that while this principle is part of the nonreductionist position, it clearly forms no part of reductionism. What matters to a reductionist like Parfit is psychological continuity; in the example above what he regards as important is that there will be someone who has his memories, beliefs, projects and so on. Since the principle that identity is what matters has been shown false by "My Division", Parfit concludes that nonreductionism is also false and that reductionism has been vindicated.

I will structure my discussion of Parfit's thought-experiment and the arguments that accompany it around a number of methodological points about thought-experiments. However, it may help to summarize here what I will argue, in the course of the discussion, about "My Division" and the debate between reductionist and nonreductionist views of personal identity. I will argue that "My Division" does not show that identity can be indeterminate; or rather, that it does so only if one assumes that identities must be grounded (in a sense to be explained). But we have good independent reasons for rejecting the view that identities must be grounded. As far as the second thrust of Parfit's argument is concerned, I will argue that even if "My Division" were successful in establishing that identity is not what matters, this would not support reductionism since the nonreductionist need not be

committed to the principle that identity is what matters. But more importantly I will argue that "My Division" does not succeed as a case against this principle. Finally I will argue that "My Division", in conjunction with another of Parfit's thought-experiments, brings to light certain internal conflicts within his reductionism. My conclusion will be that after the discussion of "My Division" we have more reason to believe in nonreductionism than in Parfit's reductionism.

SECTION 2: *The structure of revisionary thought-experiments*

In the previous chapter, we looked at one example of what thought-experiments can do. This was that they can serve to make explicit the implicit criteria we have for applying important concepts like "person" and "same person"; put more helpfully, if less carefully, they can reveal (perhaps previously unformulated) beliefs we have about persons and their identity. But something more than this is being claimed for the "fission" thought-experiments just outlined. They are being used for <u>revisionary</u> purposes: for showing that there is something wrong about our beliefs or the underlying principles we make use of in applying the concepts, and that these principles should be replaced by others. Parfit makes the following claim for "My Division": "Considering this case may

help us to decide both what we believe about ourselves, and what in fact we are" (Parfit 1984: 255 - my italics).

If there is an interesting distinction between what we believe ourselves to be and what in fact we are, then there is something immediately implausible about the use of thought-experiments towards this end, and which merits some discussion. For how can a thought-experiment reveal "what in fact we are" (as opposed to what we believe we are)? If this is what is required of a thought-experiment, is there not a real danger of attempting to infer p from ◇p? We cannot hope to discover actual contingent facts by considering merely possible worlds.

Should the proponents of revisionary thought-experiments such as the current fission one be attempting anything as rash as the route just described, they would be wrong. I am not convinced that there is not an element of that strategy in certain of Parfit's thought-experiments - in the next chapter I will discuss this at some length - but I do not think he is guilty of that mistake in this case. For although other possible worlds cannot inform us on matters of contingent fact about our own world, they are in a position to inform us on matters of possibility and necessity. For instance, the description of a possible world can show us that something presumed to be a necessary truth is nothing of the sort. And this is at least part of what Parfit attempts with "My Division"; he wants to show, among other things, that personal identity is not necessarily determinate.

To be informative, revisionary thought-experiments will have to work in this negative way; that is, by showing certain beliefs or principles to be false and thereby, perhaps, providing support for a conflicting view. In Sorensen's terms, they work as "alethic refuters", that is by showing some claim of possibility or necessity to be false (Sorensen 1992: 135)[2]. To explain how they might perform this refuting task, I will drop the "possible-worlds" talk and return to more familiar terminology. There seem to be two basic ways in which a thought-experiment could refute a theory or principle. The first way is to reveal some inconsistency in the theory itself, the second way - which is more common in the literature - is to reveal that the theory, belief, or whatever has consequences which conflict with another deeply entrenched theory or belief. Both ways thus take the form of a reductio ad absurdum.

As suggested above, "My Division" is best read as operating in this way, more particularly, in the second way mentioned. Our belief that personal identity is of necessity determinate is shown to conflict with other entrenched principles - for instance that identity is a 1:1 relation. To reject this latter principle (which is usually known as the necessity of identity) is to pay too high a cost in terms of the damage it would do to our logical system. In a similar way the belief that identity is what matters in survival is shown to conflict with the belief that the relationship between the original person and each of the survivors contains all that matters. As a result the principle of the

determinacy of identity and the principle that identity is what matters are to be rejected, and we have good reason to believe the reductionist view of persons and their identity.

SECTION 3: *Determinate identity or grounded identity?*

So far, the discussion suggests that "My Division" serves to give us interesting insights into what thought-experiments can do, in particular showing them in a revisionary role. But while it is true that we have gained insight into the workings of the method, this particular instance serves also as a guide to some important problems, for it is not as straightforward as it may seem.

Let me begin the investigation of what there is still to be learned from "My Division" by taking a critical look at the first conclusion Parfit draws from the experiment, namely that it shows that personal identity can be indeterminate. One reason for doubting that this conclusion is justified by the argument is that in order to reach it, appeal gets made to a principle which is itself questionable.

This point bears on the part of Parfit's argument in which answers (ii) and (iii) to the question "What happens to me?" are rejected. We are told that to answer that Derek Parfit survives as one of the off-shoots rather than the other

is not a viable option. The reasons provided as to why this is the case take the form of rhetorical questions:

> ...each half of my brain is exactly similar, and so, to start with, is each resulting person. Given these facts, how can I survive as only one of the two people? What can make me one of them rather than the other? (Parfit 1984: 256)

The principle giving substance to these questions is presumably that there can be no such thing as "bare identity", that there must be some discoverable property in virtue of which any claim of identity or non-identity is true, or which justifies a choice between the two options. As it is sometimes put, identity must be grounded, and without some grounding any decision that I am one or the other of the resulting people would be arbitrary.

This principle that identity must be grounded has some plausibility when expressed as it is in Parfit's questions, but when further justification is required it is not at all easy to come by. A symptom of this is the very way that Parfit makes his point: rhetorical questions take the place of any substantive argument.

That there is a problem in finding some backing for this principle can be seen elsewhere among those who are wedded to it. An important instance is Forbes, who relies heavily upon

the principle in The Metaphysics of Modality, but admits that he can see no clear way to justifying it (Forbes 1985: 127-128). This lack of positive support does not by itself undermine the principle, but things start to look worse for it when we see that there are writers who strongly deny that the grounding principle is true at all, such as Madell (1981), Garrett (1986) and Mackie (1987).[3] All three assert that there are reasons for accepting that individuals can be identical or non-identical even though there are no empirically observable facts in which this identity or non-identity consists; one strong reason will emerge in Section 4 below[4].

So far no argument has been put forward to the effect that the principle is false. That is, unless one shifts the focus and counts "My Division" as establishing the falsity of the principle that identity must be grounded rather than the falsity of the principle that identity must be determinate. This is a crucial point; for given that the status of the grounding principle is at least dubious, there is just as much reason for taking "My Division" to be a reductio of that principle as of the determinacy of identity[5]. Certainly nothing in the experiment itself suggests that only the truth of the determinacy principle is at stake. We are required to weigh one entrenched belief against another, and either can give.

This is a problem which scenarios like Parfit's, insofar as they are thought-experiments, will share with other falsifying experiments: it is not always going to be clear

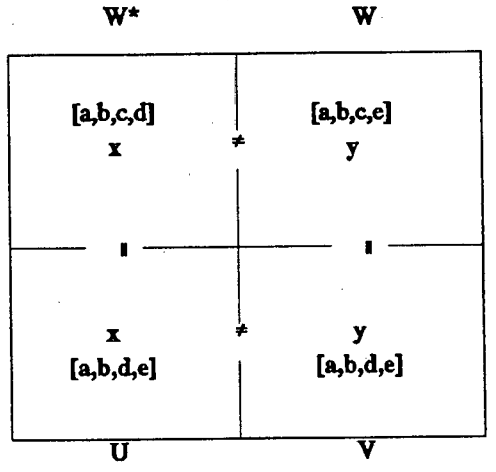
precisely what is shown to be false. This has been a standard criticism of falsificationist methodologies in the philosophy of science, but the problem becomes especially acute when the case involves one belief being bounced off other entrenched beliefs, beliefs which like those in "My Division" are metaphysical principles which are not obviously true nor necessarily clearly understood. Matters are easier when one is working within an established theory, merely seeking further support for the principles of that theory, but there is no such stable background here.

The problem need not be insurmountable, however. It may well be that other thought-experiments can be produced which also bring pressure to bear on the determinacy of identity principle without making appeal to the impossibility of bare identity. In this way, it could become evident that too high a price is being paid in clinging to the determinacy principle - that too many other principles have to be given up in order to keep that one intact. Should this be the case, it would be sound methodology to reject the determinacy principle and any other principle in a similar position. This sort of compound use of thought-experiments is an effective technique, and does occur in the literature. Examples of it, but ones which I will not go into here, are Mackie's argument mentioned above (Mackie 1987) and Parfit's own use of his "My Physics Exam" to support a conclusion drawn from his "Branch-Line Case" (Parfit 1984: 287 and 246-247)[6]. I will, however, demonstrate the technique with arguments that I offer in the following section and section 6.

SECTION 4: *An independent argument for ungrounded identity*

While this methodological point about the compound use of thought-experiments is of general significance, it should be mentioned that, even on its terms, things appear to be detrimental to Parfit rather than in his favour. For there is a thought-experiment of Salmon's, on which he bases his well-known "four-worlds" paradox, which raises a serious difficulty for the claim that bare identity is impossible (or that identity must be grounded), and without making any appeal to the determinacy of identity (Salmon 1982: Appendix 1, sec 28).

Figure 1



Assume, for ease of exposition, and since nothing crucial to the argument depends on it, that you were originally made up of four basic parts. Thus there is an individual, x, in

the actual world, w*, made up of parts a,b,c and d. Assume that there is another world, w, in which an individual y, who is not identical to x, exists, made up of parts a,b,c and e. This is not controversial: someone else could have existed who was originally made up of matter very similar, but not identical, to your own make-up. It is also hardly controversial to assume as well that you might have been made up of slightly different matter. That means that we can say that there is a world, u, in which x exists but is made up of parts a,b,d and e.

However, according to the uncontroversial assumption just invoked, there is also a world, v, in which y exists made up of a,b,d and e. Now, as a result of the transitivity and necessity of identity, y in v is not identical to x in u, since y in v is identical to y in w, while x in u is identical to x in w*, and x in w* is <u>not</u> identical to y in w. What all this means is that x in u and y in v are distinct individuals <u>even though there is nothing in which this non-identity can be grounded</u>.

So there we have independent support for the claim that identities need not be grounded, to which even one who denies that identity need be determinate must agree. As a result we have good reason to see Parfit's thought-experiment as a <u>reductio</u> of the demand for grounding rather than of the determinacy of identity thesis. It was the grounding principle that proved to be crucial to Parfit's attempt to refute nonreductionism by undermining its apparent implication that identity must be determinate. Once it is seen that his

thought-experiment assumes the truth of a contentious principle, Parfit's case for reductionism becomes less convincing than it was before. But with this independent argument against the grounding principle, "My Division" can serve as a further consideration against the grounding principle itself rather than a case against the determinacy of identity, and Parfit's case against nonreductionism appears to be anything but conclusive.

The argument against Parfit can be summed up in the following way, which may enhance its intuitive appeal. Parfit's thought-experiment really leaves us with two options: we can say that he <u>is</u> one of the survivors of the fission process (which we would like to say), or we can say that he is <u>neither</u> survivor (which we would also like to say, in order to preserve the grounding principle). Parfit in effect opts for the latter. But the argument drawn from Salmon undermines the grounding principle, so now there is less resistance to saying that Parfit <u>is</u> one survivor. Since that was one of the things we did want to say, that is what we should say.

SECTION 5: *Why the argument that identity is not what matters does not support reductionism*

In this section, in the course of raising another methodological problem facing Parfit's thought-experiment, I will argue that the second thrust of his case for reductionism

appears to be in trouble. My contention is that even if "My Division" were successful in showing that identity is not what matters, this does not imply that nonreductionism is false since accepting nonreductionism does not imply acceptance that identity is what matters. On top of this, I will argue that "My Division" is unsuccessful in its attempt to show (as a case against nonreductionism) that identity is not what matters.

Parfit suggests that "the main conclusion to be drawn (from 'My Division') is that personal identity is not what matters" (1984: 255). He presents this as being itself an argument for reductionism because he holds that only reductionism is compatible with the truth of this consequence of the thought-experiment.

"The claim that identity is not what matters," says Parfit, is "part of the Reductionist view" (1984: 264). Nonreductionism is incompatible with the consequence that identity is not what matters because it holds that persons are entities over and above any physical or psychological continuities, and Parfit suggests that implies that it is identity rather than anything else which matters in survival.

This is then Parfit's case in brief. "My Division" presents a case in which I do not survive, but in which what matters does survive: the relation between Parfit and each of the resulting people contains all that we would want in normal survival. So identity cannot be what matters. Since nonreductionists say it is what matters, they must be wrong. As a result, reductionism is vindicated.

These are the points to which Parfit makes reference when he writes that his experiment can show us "what in fact we are". It is through these points that the thought-experiment can show us that we are reducible to our bodies and/or our psychological states; as he puts it, that we can be accurately and adequately described "impersonally", that is, without any mention of our being persons (1984: 210).

I have suggested that fulfilling the aim of showing actual facts about us, despite its surface implausibility, need not be beyond the reach of thought-experiments. They can, at least in principle, reveal facts about us by showing what we are not. Thus Parfit could succeed in establishing that we are persons as envisaged by a reductionist by establishing that nonreductionism is wrong about us.

Although the strategy is acceptable in principle, I wish to make it clear that it is by no means easy for the strategy to succeed in practice. One problem we have already encountered is that the thought-experiment can backfire, and thus that the argument accompanying it needs somehow to make it plain that its target theory or principle is the only theory which can plausibly be shown false by the experiment.

However, there is another problem related to this one; for the argument also needs to ensure that the thought-experiment actually does show its target theory to be false. This is not just the previous point that something else might equally be held to have been disproved, but that whatever phenomenon is described must really conflict with the theory at stake.

The point can be illustrated with "My Division". Parfit's ultimate end is to support reductionism by destroying its rival; it is to this end that he argues that identity is not what matters. But the argument which "My Division" presents for identity not being what matters is not incompatible with nonreductionism.[7] Recall that nonreductionism is the view that personal identity cannot be reduced (without circularity) to the holding of particular physical or psychological facts; now, nothing in Parfit's argument around "My Division" prevents someone holding this and holding that identity is not what matters.

The crucial premise in Parfit's argument is that the relationship between the original person and each of the surviving people contains all that we want in normal survival. He suggests that we ought to regard the prospect of division as like ordinary survival in respect of what we want. These are claims about which attitude of the original towards the survivors is the rational one. They imply that it would be irrational to insist that identity is what a person about to split should care most about, since the relation between the original person and one of the survivors cannot plausibly be described as identity yet contains all that we want in survival. But, as suggested above, a nonreductionist can agree with all of this: nothing in her nonreductionism compels her to believe that she should not be concerned about a future individual simply because that individual will not be identical with her. What the argument shows is that to base self-interested concern about the future solely on

considerations of identity is misguided, and that considerations of the psychological connections between original and survivor are extremely important. But the thesis that persons cannot be reduced to their bodies and psychological states does not imply that we should let our concern for identity outweigh such considerations.

In this way, then, "My Division" misses its mark. This may be a problem that is peculiar to that thought-experiment, but it certainly reveals one more way in which metaphysical speculations like it can go wrong in an attempt to show our actual nature.

There is a second reason why Parfit's argument about what matters does not provide a conclusive case against nonreductionism. The reason is that from the point of view of a nonreductionist, the argument appears to beg the question. As I have outlined, the thrust of the argument is that after Parfit has divided we can no longer talk of a relation of identity between Parfit and the surviving persons, and yet the relation which does exist between Parfit and each of the survivors contains all that matters in ordinary survival. As a result, Parfit concludes that identity cannot be what matters (and thus that nonreductionism is wrong). But the nonreductionist seems entitled to balk at Parfit's assertion that identity-talk does not apply after the division takes place. It is a direct implication of nonreductionism that there can be cases in which one person is identical to another even though there is nothing (else) in which this identity consists, and likewise that there can be cases of non-identity

even though there is no observable difference to ground this lack of identity. This means that it is part of the nonreductionist's thesis that it is possible to have a case in which we cannot decide questions of identity on empirical grounds. As far as the nonreductionist is concerned, then, what Parfit has produced in "My Division" is precisely such a case - a case in which we are unable to decide on the evidence available which survivor is identical with Parfit. So far we have no case against nonreductionism. It is only if we go further and insist that neither survivor is Parfit that it follows that identity is not what matters, but to do so would be to beg the question against the nonreductionist - we would in effect be assuming that personal identity is reducible to certain observable relations. Certainly "My Division" gives us no special reason to believe that neither survivor is Parfit over and above our difficulty in deciding the issue. As a result, it appears that the nonreductionist has no reason to take the case for identity not being what matters seriously.

SECTION 6: *Another relevant example of the compound use of thought-experiments*

In discussing "My Division" and its implications for personal identity as well as methodology, it has emerged that there are a number of reasons why the thought-experiment is

not in a position to do all that Parfit claims for it. However, there is even more of relevance to our enquiry to be gleaned from this thought-experiment. I pointed out in Section 2 that thought-experiments stand to perform a revisionary function by making clearer the consequences of a theory: by showing that the theory conflicts with too many other important theories, or that it has some internal inconsistency, they can provide us with good reason for rejecting that theory.

"My Division" can serve to highlight what appears to be a significant tension of the kind envisaged in Section 2 within Parfit's own reductionist theory. At the same time, this will provide us with another example of how thought-experiments can function together to support a conclusion. The tension emerges when we consider "My Division" together with another of Parfit's thought-experiments, one which is in fact a variation of "My Division". I will argue that these two thought-experiments in combination reveal a conflict between two of Parfit's claims about what reductionism basically involves.

Parfit has claimed that a reductionist holds

> 1) that the fact of a person's identity over time just consists in the holding of certain more particular facts.

More specifically, for Parfit[8], personal identity consists in the holding of unique relations of psychological continuity - that is, x=y if and only if x is psychologically continuous with y and nobody else is (Parfit 1984: 263). To be a nonreductionist one will also hold that:

> 3) we could give a complete description of reality without mentioning persons or their identity.

Such a description would not leave anything out, even though it does not mention persons, Parfit explains, because the claim "that a particular brain and body, and a particular series of interrelated physical and mental events" <u>implies</u> the claim "that a particular person exists" (1984: 212).

My contention is that another look at "My Division" reveals a conflict between these two claims which Parfit sees at the heart of reductionism.

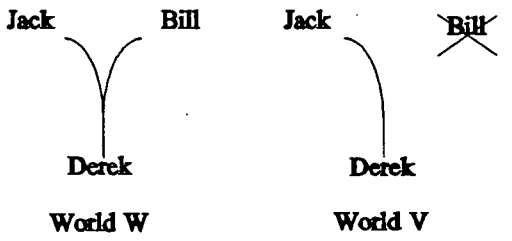
In §91 of <u>Reasons and Persons</u>, Parfit outlines the following variation of "My Division":

> I have two fatally brain-damaged brothers, Jack and Bill. A surgeon first removes and divides my brain. The halves are then taken to different wings of the hospital, where they will be transplanted into the bodies of my two brothers...Suppose that one half of

> my brain is successfully transplanted into Jack's body. Before the other half can be transplanted, it is dropped onto a concrete floor.
>
> (Parfit 1984: 269)

Parfit wishes to use this experiment to embarrass the holder of a physical criterion of identity, but let us put that aside. By itself this thought-experiment does not damage Parfit's theory, but tensions within that theory start to show when we consider this thought-experiment alongside the original version. If we call the scenario of "My Division" world "w", and this new variant "v", we get this picture:

Figure 2



Given Parfit's claim (1) and his subsequent filling-in of the details[9], in w neither Jack nor Bill has a claim to identity with Derek. This must be the case if personal identity consists in 1:1 psychological continuity, since neither Jack nor Bill has the required unique relation with Derek.

However, since there is only one survivor in v, Jack in that world can and should be described as being identical with Derek. What this implies is that using the label "Jack" in both w and v is misleading, for Jack-in-v cannot be identical to Jack-in-w. This non-identity follows, once again, from the transitivity and necessity of identity. Since Derek-in-w = Derek-in-v, and Derek-in-v = Jack-in-v, Derek-in-w = Jack-in-v. But Derek-in-w ≠ Jack-in-w, so Jack-in-w ≠ Jack-in-v.

Now, while the two Jacks are not identical, they have the same brain, the same body, and also share the series of interrelated mental and physical events. But this means precisely that the existence of a particular brain, body, etc does <u>not</u> imply the existence of a particular person: for one and the same brain, body, etc is associated with <u>distinct</u> persons in worlds w and v. In other words, assuming the theory implicit in Parfit's claim (1), his purely impersonal description of worlds w and v <u>would</u> leave something out, in defiance of claim (3) - and what the description would leave out are precisely facts about the identity of persons.

In this way, "My Division" can indeed be revealing in our enquiry. It serves as a good example of how a thought-experiment can reveal consequences of a theory which are otherwise not at all obvious, and yet are damaging for that theory. While this argument reveals tensions within Parfit's reductionism, it may not do enough damage to show nonreductionism to be true. There may be another reading of Parfit's claim (1) which does not lead to a conflict with claim (3). But even so, the argument does succeed in lending

more plausibility to the view that personal identity resists any reduction.

SECTION 7: *Conclusion*

There have been two major strands of argument in this chapter, and it may be useful to conclude by separating them out clearly so as to minimise the confusing effect this might have. The basic discussion was methodological, and here it emerged that "My Division" serves as a good example of the form a thought-experiment which stands to make us revise our views will take. Thought-experiments are an ideal way of making the consequences of a theory clear - of revealing the costs of accepting that theory in terms of its conflicts with other theories or internal inconsistencies.

However it also emerged that there are difficulties into which even as impressive and important a thought-experiment as "My Division" falls. The most significant of these is that the general problem of knowing precisely what is refuted by an experiment is particularly acute in the case of revisionary thought-experiments dealing with abstract metaphysical issues such as those at stake in the personal identity debate.

In the discussion of "My Division", what also came to light is that a thought-experiment might often function most effectively when used in combination with other thought-

experiments. Not only can this help solve the problem just mentioned, but (as I argue is the case with "My Division") this complex use of thought-experiments can also bring out consequences of a theory which would otherwise go unnoticed,

In discussing how these methodological points affect "My Division", some positive points regarding personal identity itself began to emerge. The most important is that once the method of revisionary thought-experiments is made plain, it appears that "My Division" does not offer the support Parfit envisages for his reductionism. Indeed, as a case against nonreductionism, "My Division" appears to be question-begging. On top of this, Parfit's imaginary situation when used together with the "four worlds" argument of Section 4, begins to make a case for some version of nonreductionism. This case is strengthened when "My Division" is considered alongside the alternative version of the same situation which Parfit makes use of elsewhere. In this way, "My Division" threatens to do anything but the task which Parfit sets for it. In the following chapter, it will emerge that there is further support to be found for the nonreductionist position.

NOTES

1. I stress that it is <u>according to Parfit</u> that nonreductionism implies that identity is what matters, because I will argue that nonreductionism has no such implication in section 5 below.

2. Although, as the discussion of Chapter 6 makes plain, they do not work <u>only</u> as alethic refuters as Sorensen seems to suggest (1992: 132-166).

3. Not only do these points suggest that we should be careful about a hasty acceptance of the principle that identity must be grounded, but such acceptance seems to beg the question against the nonreductionist. It is part of the nonreductionist view that I may just be one person rather than another even though there is no other fact in virtue of which this is the case. I will raise this point in more detail in Section 5.

4. One argument for ungrounded identity which has already been mentioned is Madell's thought-experiment set out in Chapter One.

5. Even though, as I later argue, the principle that identity must be grounded is false, "My Division" would not necessarily show the principle to be false as I suggest here - that is, the thought-experiment would not provide a <u>reductio</u> of the principle. It would only work as a <u>reductio</u> of the grounding principle if the principle that identity must be determinate were true. But later (Chapter 8 section 7) I will argue that Parfit is correct that identity need not be determinate. Nevertheless, his argument for this conclusion based on "My Division" still falters since it presupposes the truth of the grounding principle.

6. I discuss this case of Parfit's support of the one thought-experiment by the other in a slightly different context in Beck (1992b).

7. The argument to follow is due to Garrett (Garrett 1991: 365).

8. Although these are the "more particular facts" which Parfit favours, as I have pointed out other reductionists might point to facts other than these.

9. The argument to follow is unaffected if one favours a physical criterion rather than Parfit's psychological version in one's reductionism.

CHAPTER 8: THE TELESCOPE AND THE SPECTRUM

SECTION 1: *The fallacy of the telescope*

In the discussion of the case of "My Division", examples were mentioned of what appears to be a different kind of thought-experiment from those examined up to this point. The examples were the one which occurs in Salmon's "Four Worlds" paradox, and the one which was used to show that the existence of a particular brain, body and set of interrelated events does not imply the existence of a particular person. What initially marks these experiments as being of a different kind is that they are centrally concerned with the identity of persons <u>across worlds</u>, whereas the examples we have examined had the identity of persons <u>over time</u>, but within a single world, as their explicit concern.

This kind of thought-experiment is also of special interest to us because its use brings to the fore a methodological objection different from any raised up to this point. Whether or not the objection is relevant to thought-experiments concerned with trans-temporal identity as well is a matter that can be set aside, at least until the implications of the objection for trans-world

thought-experiments have been investigated. It is worth noting here, however, that even if the objection is not directly relevant to thought-experiments dealing with identity over time, it could still have relevance to our topic. This is because, as the discussion of the previous chapter shows, transworld thought-experiments can and do inform the debate on personal identity over time.

The objection under discussion is one put forward by Wiggins, who in turn acknowledges a debt to Kripke (Wiggins 1980: 213). Wiggins does not suggest that all thought-experiments are mistaken, but argues that "the method of possible worlds" is inappropriate in certain contexts. It would be a "mistake" to use it, he contends, in reaching conclusions regarding the necessity of identity or, what is of more direct importance to the discussion of the identity of persons, the necessity of origin.

The method may be useful in investigating identity over time - indeed, Wiggins himself makes such use of imaginary cases in his Identity and Spatio-Temporal Continuity (Wiggins 1967, part 4.3) - but here it reaches a limit. The reason for the method's being inappropriate with regard to the necessity of origin debate, according to Wiggins, stems from the nature of possible worlds. The misgivings emerging in his mind will be set out below. It is important to note first that, while the discussion to follow is aimed only at the necessity of origin, similar considerations could be produced which would affect other transworld relations between individuals or transworld identity principles, and thus that the objection

has wider significance than it may seem to have at first glance.

The thesis which bears the label of "the necessity of origin" is this: the identity of an individual depends in a crucial way on that individual's origin; a different origin would entail a different individual. This plant could not have grown from a seed other than its actual parent seed, nor could this child have had different parents[1]. Wiggins's contention, then, is that the truth of these claims cannot be supported by argument from possible worlds.

Wiggins sees the believer in the necessity of origin postulating worlds different from the actual one in that some individual has an alternative origin, and then asking whether the resultant plant, person, or whatever is identical to the corresponding individual in the actual world. The supporter of the necessity of origin concludes, from due consideration of the alternative world, that there is no relation of identity between the individuals concerned. Conversely, the opponent of the necessity of origin may postulate a similar counterfactual situation and conclude that the individuals are identical.

Wiggins suggests that this method of reasoning is faulty as a result of the conception of possible worlds which underlies it. For possible worlds, following Kripke, are suppositions, scenarios which we construct: "a possible world is given by the descriptive conditions we associate with it" (Kripke 1980: 44)[2]. For such a scenario to be complete, its supposer is required to stipulate the identities of the

individuals which appear in it; if we are to be able to identify the relevant possible world, then we must already know the identities of its component individuals.

In this way, questions of identity must be decided in the postulation of possible worlds. Precisely what one cannot do is to examine a world one has constructed and "read off" the identities of the individuals appearing there. Kripke warns against a "distant countries" conception of possible worlds - they are not like faraway countries or planets to be discovered and investigated with the use of powerful telescopes (Kripke 1980: 44). Wiggins accuses those who use possible worlds to argue for (or against) the necessity of origin of being guilty of treating them in just this unacceptable manner - of committing what I will call the fallacy of the telescope. It seems then that one of the limits on the use of possible worlds is that they cannot decide the question of the necessity of origin.

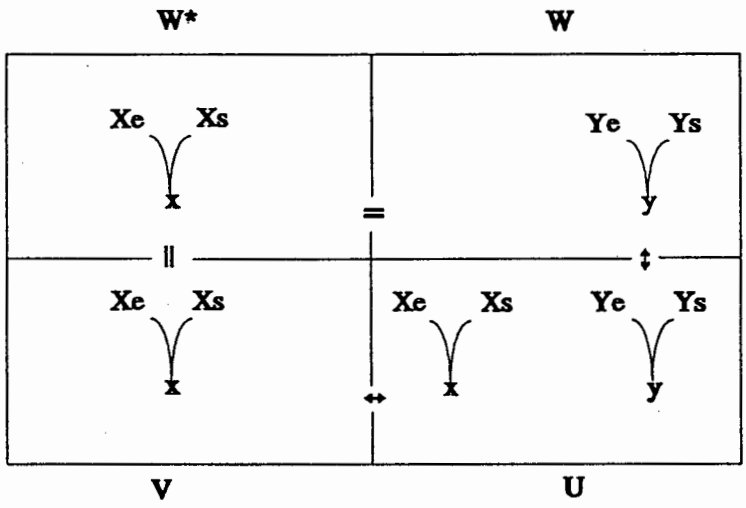
This limit deriving from Kripke and Wiggins is a false limit, however. Or rather, Wiggins does not succeed in establishing that the use of possible worlds is limited in the way he contends it is, even if he is right about the nature of possible worlds. By way of demonstrating this, I propose to take a look at an important argument in favour of the necessity of origin which depends crucially on an appeal to other possible worlds, and argue that it is innocent of the telescope fallacy.

SECTION 2: *A counter-example to Wiggins on the telescope*

The possible-worlds argument for the necessity of origin alluded to above is that of Forbes (Forbes 1985: Ch 6 sec 3). As he presents it, it concerns oak trees and acorns, but it is intended to apply to other kinds of individuals as well, persons included. I will recast the argument in terms more directly relevant to my topic.

The argument in favour of the necessity of origin works as a reductio ad absurdum of its denial. The crux of the case is that if one rejects the necessity of origin thesis, one commits oneself to certain unacceptable ungrounded identity claims. As a result, scepticism about the thesis should be rejected. Suppose that in the actual world w*, x is the child of Xe and Xs. If the thesis were false, then in another world, w, the same child could have been born to Ye and Ys.

Figure 3



('=' symbolises identity, '↔' indistinguishability)

But then consider a world, u, which is just like w - there is a child born to Ye and Ys in u which is indistinguishable from that in w - except in that there is also a child born to the X's. Which child in u is identical to the actual one? It can't be the x-child, because then the y-child in u would not be identical to the y-child in w, since the latter <u>is</u> the actual child. The trouble is that there is no relevant difference between the y-child in u and the y-child in w in virtue of which this non-identity could obtain[3]. Since the two are indistinguishable in all relevant respects - that another child also happens to exist in one of the worlds does not seem to count as a relevant respect - the sceptic would be left with an ungrounded non-identity claim. The sceptic must then (given the formal properties of identity) say either that the y-child in u is the actual child, or that neither child in u is the actual one.

Whichever he chooses, it follows that the x-child in u is not the actual child. But then consider a fourth world, v, in which a child just like the x-child in u exists. Since it is quite plausible that the actual child could have been just like the x-child in u, we can claim that this <u>is</u> the actual child. It follows, then, that the x-child in v is not identical to the x-child in u, since according to the sceptic the latter is not the actual child, whereas the former is. The two are indistinguishable in all relevant respects, however, and so the sceptic is once again faced with an ungrounded claim. Since all the sceptic's options have been

used up, the conclusion must be that scepticism about the necessity of origin is mistaken.

That rather complex case is Forbes's argument for the necessity of origin. It is a valid argument even if it contains some contentious assumptions, to which I will return below. But, and this is where it is relevant to our present concerns, nowhere does Forbes attempt to, or need to, "read off" the identities of individuals from a given possible world. He does not claim to discover any identities or non-identities in the way envisaged by Wiggins. His argument may hit problems somewhere along the line - indeed, we have seen strong reasons for rejecting his assumption that identities must be grounded - but there is nothing wrong with its use of possible worlds. In the rest of this section, I will elucidate this defence.

Forbes's argument turns on the similarity in all relevant respects between individuals in various possible worlds. He argues that if one denies the necessity of origin, then one is inevitably left with identity claims which are arbitrary or ungrounded; specifically with assertions denying identity between individuals in the absence of any relevant difference between those individuals. This may still sound a bit dubious given Wiggins's misgivings: Wiggins's view suggests that if the identity of any individual is not made plain in the postulation of a world, then the world is incompletely specified, and thus unacceptable in an argument. But this is not true of Forbes's worlds. Forbes is not guilty of making unacceptable use of incompletely specified worlds. He

circumvents this charge by arguing in effect that no matter what identity you assign the individuals in his set of worlds, if you don't accept the necessity of origin, certain of the identities you assign will be ungrounded. Thus Forbes neither deals with incompletely specified worlds, nor does he commit the telescope fallacy - and yet he presents a case for the necessity of origin which is well worth consideration.

The conclusion to be drawn is that it has not been established that arguments from possible worlds to the necessity of origin involve a defective conception of what possible worlds are. There is some importance to this conclusion. Wiggins ends up agnostic about the necessity of origin (Wiggins 1980: 213) because he sees possible worlds arguments as being inappropriate, and yet the only kind of argument which could decide the issue. The case just presented shows that there can at least be sensible debate on the topic. And since Wiggins's argument fails against its direct target, namely, the necessity of origin, we also have no reason to believe that arguments concerning other transworld identity principles will inevitably commit the telescope fallacy.

SECTION 3: *Trans-temporal identity, the telescope and the Combined Spectrum*

So far I have argued that possible worlds are not by their very nature prevented from contributing to conclusions regarding the necessity of origin or other principles of transworld identity, but the question remains open as to whether arguments committing the telescope fallacy occur in discussions of the conditions of personal identity over time.

In this section I wish to examine a centrally important argument in the debate between reductionist and nonreductionist views on personal identity. This argument relies heavily upon the use of possible worlds (although it is not phrased in possible-world terms). I will argue that it is unsuitable as support for the conclusions drawn, and that this infelicity is a result of, amongst other things, the abuse of possible worlds. At least on one plausible interpretation of the argument, a mistake closely akin to the telescope fallacy is made. I will set out the argument more or less as it is presented in the literature, and then in the following sections show how it is problematic.

The argument concerned is one of Parfit's, which he calls the "Combined Spectrum" (Parfit 1984: 236-244). The spectrum consists of a range of possible cases in which a person undergoes increasing psychological and physical changes along the range. Taking the person concerned to be Parfit himself, each case is a possible world in which Parfit is operated upon and physically altered to an increasingly greater degree. In the first case nothing at all is changed. In the second case

a few of his body and brain cells are changed, replaced by new ones. Since our psychological features are dependent on our brain states, the result of this change will be certain slight psychological changes. The new cells which replace Parfit's old cells will be replicas of the cells of another person - Parfit selects Greta Garbo as his model. Along the spectrum, increasing numbers of cells are replaced accompanied by increasing psychological changes - Parfit will gradually acquire some of Garbo's beliefs, etc - until in the final case the result of the operation is an exact replica of Greta Garbo - exact both physically and psychologically.

Each of these worlds is indeed possible, according to Parfit; at the moment they are merely technically impossible. He holds that the only crucial assumption is that our psychological states depend on our brain states, and nobody is about to deny that. From consideration of this range of possible cases Parfit draws the conclusion that "these cases provide, I believe, a strong argument for the reductionist view" (Parfit 1984: 237).

Reductionists like Parfit believe that the relation of personal identity can be analysed into more familiar relations, such as those of psychological or physical continuity. They believe further that these relations can be described in purely impersonal terms; it is this which really makes them reductionists. Nonreductionists deny that any such analysis is adequate, or deny that the reduction to impersonal terms goes through. Another important difference, according to Parfit[4], is that nonreductionists believe that our identity

is an all-or-nothing matter: one either is or is not identical to a given person; there are no indeterminate cases.

This is where the Combined Spectrum comes in. Parfit contends that the identity of the person after the operation, while clear in the cases at either end of the spectrum, is indeterminate near the centre of the spectrum. The question, "Do I continue to exist?" is in these cases an empty question. If an all-or-nothing view of identity were true, there would have to be a sudden change of identity between two of these central cases; the resulting person in one being Parfit, and in the next one Garbo. As it is there is no sudden psychological change marking this change of identity to be observed. All that occurs is the change of a few brain cells accompanied by a slight psychological change, and one could hardly base a claim about something so momentous as change of identity on a change so insignificant.

The nonreductionist, however, is bound to the unacceptable claim that there is a sudden change of identity somewhere along the spectrum, although we could never know where on the spectrum it occurred. Since this is an untenable position, the Combined Spectrum shows nonreductionism to be false and supports the reductionist view.

SECTION 4: *Why the Combined Spectrum does not support reductionism*

I wish to argue that the case which the Combined Spectrum provides for reductionism is not nearly as strong as Parfit suggests. As a first point I will argue in this section that Parfit's thought-experiment only raises problems for some nonreductionists, and not for nonreductionism as such. If this contention is correct it is an important one, since Parfit's case works as a revisionary thought-experiment in the way outlined in Chapter 7. That is, he intends it to support reductionism by showing nonreductionism to be false. If, however, only a specific type of nonreductionism is shown to be false then the case would provide no more support for reductionism than for other types of nonreductionism.

The argument of the Combined Spectrum turns on the issue of the determinacy of identity. What emerges from it is that if one insists that personal identity must be determinate, then one is forced to face the unpalatable consequences Parfit outlines, that somewhere on the spectrum there is a sudden change of identity even though we can never know where that change occurs. Parfit sees this as providing support for reductionism because (as outlined in Chapter 7) he takes the nonreductionist to be tied to the principle that identity must be determinate. I will argue that the Combined Spectrum misses its mark because a nonreductionist can consistently reject that principle.

Nonreductionists have tended to accept that identity must be determinate: prominent examples are Butler, Reid and

Swinburne. It is not hard to see why a nonreductionist might see this principle as part of their nonreductionism. The reason lies in the Cartesian view of a person which many nonreductionists share. That is, they see a person as an immaterial entity - a Cartesian ego - that exists independently of any particular physical or psychological features that person may happen to have. Physical and psychological facts thus appear to be irrelevant to personal identity; identity becomes a matter of being the same immaterial entity, not of having the same body or the same memories and beliefs. As long as this is the case, there seems to be no room for indeterminacy with regard to questions of identity. It would only be once one acknowledged the importance of facts like psychological or physical connections which are matters of degree that there would be room for indeterminacy regarding questions of identity. As a result, Parfit takes it that reductionists can accept the possibility of indeterminate identity while nonreductionists cannot.

However, it is just not the case that the sort of nonreductionism which is Parfit's explicit target is incompatible with indeterminacy. To recall, the nonreductionist whom Parfit is attacking is anyone who believes that personal identity does not just consist in the holding of physical and/or psychological facts or who holds that these facts cannot be impersonally described (Parfit 1984: 210-212). The latter part of this description is telling, for it acknowledges that a nonreductionist may see physical and psychological facts - the very things which

introduce indeterminacy to questions of identity - as an important part of the story. This nonreductionist will insist that these facts ultimately presuppose personal identity in some way, and thus cannot be "impersonally described", but that is not in any way to mark them as irrelevant to the concept of personal identity.

A nonreductionist who takes this line, while being true to the spirit of the position which Parfit is opposing, will not share the Cartesian view of persons outlined above. Such a nonreductionist would deny that persons are separately existing immaterial entities. It may help to make the distinction between the two kinds of nonreductionism clearer to appeal to the distinction Wiggins draws between the "is" of identity and the "is" of constitution (Wiggins 1980: 30). The "is" of constitution is the sense used in saying that a souffle is simply eggs and milk, and it does not have the same sense as the "is" in "a souffle is a souffle". The Cartesian nonreductionist holds that a person is not just a brain, body or set of interrelated physical and psychological events in either sense. The non-Cartesian nonreductionist holds that while a person is not just a brain, body or set of interrelated psychological and physical events in the sense of identity, a person is no more than that in the sense of constitution.

My suggestion here is that one may well deny that personal identity can be successfully reduced to (say) psychological continuity, while still holding that the concept of psychological continuity is central to the analysis of our

concept of personal identity. The Combined Spectrum would then only affect a nonreductionist who - on top of his nonreductionism - also sees particular psychological and physical facts as irrelevant or who for some independent reason accepts that identity must be determinate, but it would not affect nonreductionism as such. It would be a threat to the Cartesian nonreductionist, but not to a nonreductionist who rejects the view of persons as separately existing entities. To put the point in slightly different terms, what the Combined Spectrum stands to show is that the concept of personal identity is analysable, but not that it is reducible.

Since the terms "analysable" and "reducible" have often been treated as synonymous in the literature, the distinction drawn here between the two needs some discussion. The question is bound to arise as to whether there is any point in asserting the analysability while denying the reducibility of personal identity, even if one grants that it is coherent to do so. If personal identity is not reducible to psychological continuity, then what status does an analysis of the relation in terms of psychological continuity have? Why should psychological matters be of any concern to the nonreductionist? The answer to these questions begins with the point that not all good analysis takes the form of reduction.

An example will help. Locke's analysis of personal identity, outlined in Chapter 1, has famously been criticized by Butler as being circular. Locke suggests an analysis in terms of memory:

> as far as any intelligent being can repeat the idea of any past action with the same consciousness it had of it at first, and with the same consciousness it has of any present action; so far is it the same personal self. (Locke 1694: II,xvii,10)

Butler's response is that memory presupposes personal identity, just as knowledge presupposes truth, and thus the analysans only works by having the analysandum contained in it (Butler 1736: 124).

Butler seems to be right. You can only remember your own experiences - were they not yours, what was going on would fall under imagination or telepathy, but not memory. Since the analysis is circular, Locke cannot be said to have offered a reduction of personal identity. But while this is true, Locke's analysis is not circular in a way which renders it trivial and devoid of insight.

For although the ability to remember may presuppose identity, memory is a concept distinct from personal identity. What Locke's analysis reveals and investigates is an extremely important connection between the two concepts, even if that relationship is not reductive. The analysis serves to make clearer the position these two concepts have relative to each other in our folk-psychological conceptual scheme. This sort of analysis will be especially important if it reveals links between the concept under discussion and another concept with

a well-established place in that scheme - a concept with which we are more familiar and understand more clearly[5].

The arguments of Chapter 7 in support of a nonreductionist view suggest that any analysis of personal identity is bound to be somehow circular, in that at some stage the notion of personal identity itself will be used in the analysis. This suggestion gets some support when one looks at the detail of a reductionist account. Parfit's reduction of personal identity is to the effect that "Personal identity consists in (psychological connectedness and/or continuity, with the right kind of cause) holding uniquely - holding between one present person and only one future person" (Parfit 1984: 263). Quite clearly, the relation into which personal identity is apparently being reduced contains reference to persons, and thus indirectly to personal identity itself[6]. But this by no means implies that non-reductive analysis is unimportant or that we should give up the attempt at establishing connections between personal identity and other concepts. Since there is an important difference between analysability and reducibility, and since one can sensibly accept the analysibility of personal identity while denying its reducibility, Parfit's case against that relation's being unanalysable does not imply it is reducible.

Having made these points about the difference between analysability and reducibility and shown that the Combined Spectrum does not affect a sensible variety of nonreductionism, I will for the sake of argument ignore the differences within the nonreductionist camp, and treat

nonreductionism as a single doctrine - the Cartesian variety which includes adherence to the determinacy of identity - as Parfit does in his exposition of the Combined Spectrum.

SECTION 5: *One interpretation of the Combined Spectrum*

It has just been argued that Parfit's Combined Spectrum argument misses its target of nonreductionism, and thus does not provide the strong argument for reductionism which Parfit wanted. The charge laid against it earlier, however, was a much more damaging one, namely that it committed a version of the telescope fallacy. As was mentioned, this charge is not a straightforward one, and needs further explanation. There are two plausible interpretations of Parfit's argument, and only one of these involves the fallacy. As I will point out, Parfit makes use of elements of both interpretations in his discussion without making a clear distinction between the two, and this creates tensions. Much of what Parfit stresses in his discussion points towards the primacy of the interpretation which does involve the fallacy, which runs as follows.

The Combined Spectrum is claimed to support reductionism by showing that our identity can be indeterminate. It shows such indeterminacy by establishing certain facts about us. The indeterminacy follows once it is established that we (i.e.

persons) are complex entities in a special sense. That is, that we are entities made up of psychological and physical components, components such as beliefs and desires (amongst many other things) which can be atomistically separated out, removed, replaced and re-assembled without the functioning of the whole system being affected in any radical way. If persons are in fact such entities, then we must acknowledge that their identity can be indeterminate, as in the central cases of the spectrum. I have no wish to argue with the view that personal identity can be indeterminate; the arguments of sections 3 and 4 of Chapter 7 were in defence of the possibility of ungrounded identity, and not against indeterminate identity. Indeed, in section 7 below I will argue that there are familiar and not at all fantastic examples of a very similar sort of indeterminacy to be found.

But this agreement does not preclude a question of principle being raised against Parfit's particular argument. For Parfit's spectrum needs to <u>show</u> that we are in fact complex in the sense outlined, and I wish to deny that a series of possible worlds - a set of <u>suppositions</u> - can establish such facts about the actual world.

The proposed interpretation of the argument of the Combined Spectrum is that the thought-experiment establishes that our identity can be indeterminate by showing that we are complex entities. This interpretation is a plausible one given Parfit's stress on the importance of <u>evidence</u> to his case. The conclusion of his discussion of the Spectrum is that what he has done is to marshal overwhelming evidence

against nonreductionism. He ends off: "There is no evidence in favour of this view [viz. nonreductionism], and overwhelming evidence against it" (1984: 243).

I will argue that the Combined Spectrum can only provide evidence (in Parfit's own sense) against nonreductionism by committing the telescope fallacy. To get an idea of what he has in mind when he talks of evidence and what evidence he actually produces against nonreductionism we can consider the following passage from his discussion of the Spectrum.

> Except for the cases close to the near end, the cases in the Combined Spectrum are, and are likely to remain, technically impossible. We cannot therefore directly discover whether the results would be as I assume...But what the results would be depends on what the relation is between the states of someone's brain and this person's mental life. Have we evidence to believe that psychological continuity depends chiefly, not on the continuity of the brain, but on the continuity of some other entity, which either exists unimpaired, or does not exist at all? We do not in fact have the kind of evidence that I described above. And we do have much evidence both to believe that the carrier of psychological continuity is the brain, and to believe that psychological connectedness can hold to any reduced degree. (Parfit 1984: 238)

Parfit's crucial evidence against nonreductionism is that mentioned in this passage - evidence that our psychological states are dependant on our brain states. But this evidence provides no direct support for the view that personal identity

can be indeterminate[7]. A Cartesian can accept that dependency at no risk of being inconsistent. Perhaps Parfit's claim that the evidence supports reductionism is misleading and what he means is that the facts about psychological dependency are evidence for the plausibility of his thought-experiment which is in turn a kind of evidence (in the much weaker sense of showing it to be a coherent thesis) for reductionism. The fact that our psychological states are dependant on our brain states would then contribute by providing evidence that we are in fact complex entities.

But that fact provides no such evidence: the dependence of our psychological states on our brain states implies nothing about our complexity. It would be absurd to deny that we are complex psychophysical entities in one ordinary sense of "complex" - it has obviously been established that we are far more intricate systems than amoeba or fish. But complexity in the required sense of having the capacity to be taken apart belief by belief and then reconstructed and altered does not follow from the premise for which Parfit has support, that our mental states are a function of our physical ones.

So the evidence Parfit marshals prior to his description of the Spectrum is neither direct evidence for reductionism, nor evidence for our complexity. But then perhaps complexity is something he wishes to assume in the argument for indeterminacy. However, this option is not plausible. Not only would it make his stressing of the dependence of psychological states on physical ones totally irrelevant, but

complexity is not something Parfit can simply assume. He cannot do so because there are strong reasons for denying that persons can be taken apart and reconstructed piece by piece. On the contrary, the holistic character of our mental states like beliefs and desires seems undeniable.

As part of the system of folk-psychology outlined in Chapter 2, particular beliefs and desires function to explain behaviour only against the background of a vast network of other beliefs, desires, and so on. Even though one wants some rhubarb tart and believes that the tart on offer is a rhubarb tart, one will not eat tart if one believes that it is poisoned. And even if one does not believe it poisoned, the tart will remain uneaten if one has formed the intention never to accept anything from the person offering it, and so it goes on. What follows is that we have no reason whatsoever to believe that a single mental state, or even a set of mental states, could be removed from an individual's psychology leaving the remainder intact.

To put it in less abstract terms, we have no reason to suppose that, for instance, Parfit's belief that personal identity is indeterminate could be transplanted in isolation into the brain of a blonde actress and remain that belief, suddenly appearing among whatever else she may happen to believe. Given this holistic picture, it appears that the outcomes which Parfit envisages in his scenarios may well be impossible in a far stronger sense than that of technical impossibility. Any assumption about our atomistic complexity

is totally at odds with the holistic scheme of folk-psychology in which beliefs and desires find their place.

Nor should it be thought that these difficulties depend upon the truth of Davidson's anomalous monism (Davidson 1970), which made these claims about mental holism prominent. For Parfit's picture is unlikely to be acceptable even to those who accept the possibility of psycho-physical reduction. Even philosophers who accept a form of type physicalism see that it is vain to hope for the establishment of psycho-physical identities other than relative to a particular individual at a particular time (cf Loar 1981). In the light of this, there is no reason to believe that the transplanting of a small amount of brain matter from one individual to another will bring with it a particular mental state.

But if Parfit does not, and indeed may not, make the assumption that we are complex entities, and the facts which Parfit produces provide no support for that claim, then it must be that the Combined Spectrum itself somehow establishes our complexity and in <u>that</u> way provides evidence for reductionism, or the argument fails. That brings the problem to a head. The Combined Spectrum is obliged to establish that we are in fact complex. If we are indeed complex then that is a contingent fact about us, and to reach such a fact on the grounds of the series of suppositions which we are offered, the telescope fallacy must be committed. If Parfit wants his Spectrum to provide evidence for reductionism, as he claims it does, then this outcome is inevitable.

SECTION 6: *An alternative interpretation*

There is another interpretation of the argument which is also plausible, and which allows it to avoid the fallacy. While this may suggest that this interpretation is thus to be preferred, it is nevertheless the case that this alternative reading is itself problematic. On top of this, it makes no sense of Parfit's insistence on the important role of evidence in his argument.

The alternative interpretation, put briefly, is this. Nonreductionism holds that personal identity is always determinate. The Combined Spectrum sets up a scenario in which identity is <u>not</u> determinate. So nonreductionism is false. All that is required of the Spectrum is that it demonstrate the possibility of the operations having the relevant outcomes; if they could occur, then identity could be indeterminate. Thus, as long as there <u>could</u> be persons who were complex in the required sense, the argument would go through. In this way, it makes no difference as to what is actually the case about us persons in this world, and so there is no threat of the fallacy being committed. One might also note that it is irrelevant to this argument that our psychological states are actually dependant on our brain states; the mere possibility is enough for the argument's purposes, thus leaving much of Parfit's exposition rather puzzling.

Apart from the uneasy fit of this interpretation with parts of Parfit's exposition, there is also room for the nonreductionist to respond, even a nonreductionist of the Cartesian variety.[8]

The most effective response a nonreductionist can make is to bring up the holism of the mental once again. In the previous section, appeal was made to mental holism in order to show that Parfit could not simply assume that persons are complex in the sense that they could be taken apart piece by piece and reconstructed or changed. But the same considerations have the further implication that Parfit's Spectrum argument, even interpreted in the current way, collapses.

Parfit's setting-up of the argument requires that we answer his question after each operation in the spectrum, "Would the resulting person be me?" His charge is that in relation to the cases around the middle of the spectrum the question would be an empty one, in the sense that either a "yes" or a "no" answer would be as good as the other. But it is quite clear that while Parfit believes the answer to his question may be indeterminate, he believes that his question makes sense; that is, he takes it that what will emerge from the operation in these cases will be a person.

It is here that mental holism becomes important - because, given the extraordinarily complicated interdependency of a person's mental states, we have good reason to believe that whatever emerged, it would not be recognisable as a person at all. To add, for example, one belief to a person's

psychological system can have repercussions throughout the system in terms of changed beliefs, desires, fears, behavioural tendencies, and so on. To add, remove and change vast numbers of mental states as Parfit proposes would inevitably have repercussions of an enormously greater magnitude. And we still need to add to this picture that the states to be added come from another equally complicated independent system.

Now, to suggest that out of the clash of one half of each of these systems with one half of the other one coherent mind will emerge is almost absurdly hopeful, and is something we have no reason at all to expect. Yet that is precisely what we have just seen that Parfit does require for his argument to have any intuitive force. Because the Combined Spectrum, even interpreted in the second way, has this unacceptable requirement, the nonreductionist is entitled to take his view to remain intact after Parfit's challenge.

SECTION 7: *An independent argument for indeterminacy*

One final point remains to be made before we leave the discussion relating to the determinacy of identity. I have argued in Section 4 that nonreductionism of a non-Cartesian variety, which accepts the importance of psychological continuity and denies that persons are separately existing

entities is compatible with indeterminacy. I have nevertheless also argued, in Chapter 7 and the previous two sections, that Parfit's influential arguments fail to establish that personal identity can be indeterminate. This makes it seem that a stronger form of nonreductionism than the kind favoured by the argument of Section 4 remains a viable option, and that the compatibility of the non-Cartesian variety with indeterminacy is not necessarily a reason for adopting that variety of nonreductionism. The final point I wish to make is that this compatibility is a definite plus, because the possibility of indeterminacy can be established by arguments other than Parfit's.

It is not necessary to go to the fantastic lengths which Parfit does to achieve his conclusions with regard to determinacy and the identity of persons; indeed, the arguments of this chapter show that it is also unwise to go to such lengths. The fact that personal identity can be indeterminate emerges from contexts which are more familiar and less far-fetched.

Consider, for instance, the case of an individual who suffers from Huntington's chorea. The results of this disease are physical deterioration and gradual psychological deterioration. Eventually a point is reached when the individual is no longer recognisably a person at all. Since the deterioration is a gradual one, there will be some point during the disease's development when it will be indeterminate whether we are still confronted by a person or not. The same will be true in the cases of certain sufferers from

Alzheimer's disease. Now while the indeterminacy in these sorts of case is not strictly the indeterminacy of identity which Parfit is attempting to illustrate, it is nevertheless an indeterminacy which must be extremely embarrassing to the nonreductionist who believes that identity must be determinate. And we can link it to the kind of indeterminacy which concerns Parfit. The Cartesian, adhering to the determinacy of identity, is obliged to say that at some particular stage the sufferer ceases to be a person or ceases to be who he was, even though we can never know which stage this is; and that it is precisely the kind of unacceptable position into which Parfit wants to push the nonreductionist. Since the possibility of indeterminacy is one which we must take seriously, then, the weaker form of nonreductionism advocated above is definitely the form to be preferred.

SECTION 8: *Conclusion*

The upshot of this discussion of the two interpretations of the Combined Spectrum is that the telescope fallacy is indeed a factor in one of the most important and influential thought-experiments in the literature. For points which Parfit stresses suggest that the Spectrum is to be interpreted as an argument which involves the fallacy; and that

interpretation is only avoided by invoking another which also does not refute the nonreductionist.

What this means as far as methodology is concerned is important. It suggests that in using thought-experiments we must bear in mind that we are dealing with possible worlds, and that subject matter demands special care. Specifically, even though an experiment is not expressed in possible-world terms, we must be careful not to try and read off too much from it. Thought-experiments can reveal things about our beliefs and their relative strength and importance, and they can also inform us that certain beliefs or principles are incoherent or in conflict with some fundamental belief; but they cannot get past this to reveal actual contingent facts.

As far as personal identity itself is concerned, the discussion of this chapter shows that Parfit's Combined Spectrum does not make the strong case for reductionism which he wishes it to make. It does not present a convincing case against a non-Cartesian version of nonreductionism, nor does it even succeed in refuting the much stronger Cartesian version.

Nevertheless, one central point which Parfit wishes to establish, namely that personal identity can be indeterminate, is one which can be established by considerations other than the Combined Spectrum. What this means is that nonreductionism incompatible with such indeterminacy is to be rejected. But this is no victory for reductionism, since in Section 4 I pointed out how a nonreductionist can hold a position which is compatible with indeterminacy by allowing

that personal identity is to be analysed in terms of psychological continuity while denying that it can be reduced to such terms. As a result, the arguments of this chapter both defend nonreductionism against Parfit's attack, and give us further reason to be sympathetic to a version of nonreductionism.

NOTES

1. Strictly speaking, the doctrine concerns particular egg and sperm cells, and not parents in the usual sense.

2. There are other views as to what possible worlds are, and these may well have different consequences from Kripke's view. Kripke's view is nevertheless widely enough accepted for us to concentrate on it here. I investigate some consequences of other views in my paper, "The Method of Possible Worlds" (Beck 1992b).

3. Importantly, Forbes excludes the properties of other individuals from counting as "relevant differences". Identities must, he feels, be intrinsically and not extrinsically grounded (1985: Ch 6 §4).

4. Again I stress that this is according to Parfit because I will deny the truth of this claim in what follows.

5. These points are endorsed by Wiggins: "No reduction of [sameness or identity] has ever succeeded...Nor is it called for, once we realize how much can be achieved in philosophy by means of elucidations which use a concept without attempting to reduce it, and, in using the concept, exhibit the connexions of the concept with other concepts that are established, genuine collateral and independently intelligible" (Wiggins 1980: 4).

6. Garrett makes this point in more general terms.

> In each version (of reductionism) there is reference to persons in the analysans (e.g., in the RHS (right-hand side) of the Psychological Criterion there is reference to persons A and B), and the concept person, in virtue of being a sortal concept, 'contains' the criteria of personal identity over time, precisely the criteria we are attempting to elucidate.
>
> (Garrett 1991: 362)

7. As Madell has pointed out in "Derek Parfit and Greta Garbo" (Madell 1985).

8. In "Finding Ourselves: Personal Identity and the Limits of Possible-World Arguments" (Beck 1991), I argued that nonreductionists can defend their position against the Spectrum with two arguments. Firstly that a Cartesian nonreductionist can claim that while we are separately existing entities, we might not have been, and secondly that the indeterminacy which the Spectrum reveals is only an epistemological one; that is, that there is a fact of the matter who emerges in the central cases, it is just that we don't know who it is. While there may be something to both of these arguments, I feel now that this response is just too heroic to be really plausible - no real nonreductionists put forward either argument. The argument below serves far better to show why the Spectrum poses no threat to nonreductionism.

CHAPTER 9: CONCLUSION

SECTION 1: *What the thesis has achieved*

The arguments of Part One led to the conclusion that the current unpopularity of thought-experiments is undeserved, since the considerations on which it is based are misguided. Flew's attack on thought-experiment rested on a misconceived picture of how terms like "person" and "same person" get their meaning. Fodor appears to run together the question of what we would say if things were different given our conceptual scheme with the question of what our conceptual scheme would be if things were different, and this casts doubt on his conclusions about the acceptability of thought-experiments. Wiggins's attack on fission thought-experiments involved treating person as a natural kind, but good reasons for not doing so emerged. Revising Wiggins's position in the way proposed by Kitcher and treating person as a law-governed kind rather than a natural kind offered a preferable alternative, but failed to rule out the sort of thought-experiment in question. I also argued that Wilkes's attempt to show fission and other thought-experiments to be suspect were inconclusive.

Having cleared the method of thought-experiment of its bad name, the discussion of Part Two revealed something of what the method is capable of showing in the context of personal identity. One important function thought-experiments can perform is to reveal the relative importance of the principles of classification which are implicit in our use of the concept of being the same person, by drawing out our intuitions as to when the concept does or does not apply. Thought-experiments can reveal which implicit principles we are least prepared to give up or deny. Used in this way, they can serve as a way of supporting a theory of personal identity, and as a way of attacking theories as well. They can support a theory by showing its principles to be in accord with our intuitions as to when the relevant concept applies. They can provide ammunition against a theory by showing its principles to be at odds with such intuitions. In line with this, I argued that it emerges from thought-experiments that considerations of psychological continuity are more fundamental to making judgements about the identity of persons than those of physical continuity.

Thought-experiments can also be used against views in the personal identity debate by showing that the view in question suffers from internal inconsistencies, or showing that it has consequences the costs of which are not worth paying. Such costs are incurred, for instance, when a thought-experiment reveals that the view conflicts with some fundamental metaphysical principle like the transitivity of identity. This aspect of the method of thought-experiment can also be

used to support a view in the debate by showing its opposition to suffer from one or other of these disorders. Parfit's case of "My Division" provided a clear example of an attempt at using a thought-experiment in a supporting role in this way. Although it provided a clear example, closer investigation suggested that it did not provide an example of a revisionary thought-experiment which actually achieves its aim. What emerged from the thought-experiment and ensuing discussion was not the revisionary reductionist view Parfit favours, but that when used in conjunction with other thought-experiments "My Division" gives support to a nonreductionist view opposed to Parfit's own.

Using thought-experiments in these various ways, we saw that support emerged for a particular view of personal identity which is nonreductionist but also non-Cartesian. It is non-Cartesian in that it denies that persons are immaterial entities existing apart from any physical or psychological facts. Persons are not their bodies, or their experiences; but nor are they immaterial entities. In this way the view is neither incompatible with materialism nor with the possibility of identity being indeterminate. The theory is nonreductionist in that it denies that personal identity can be reduced to impersonal terms. It allows that personal identity can be analysed in terms of psychological continuity, but denies the success of any further attempt to reduce our identity to psychological continuity or to describe the analysans without mention of persons and their identity.

A view along these lines begins to emerge when the method of thought-experiment is taken seriously. In the course of this discovery, a number of important methodological points regarding thought-experiments came to the fore. We cannot expect thought-experiments to provide a direct route to metaphysical truth. We can neither expect them to reveal ungainsayable intuitions - absolute truths - nor to provide evidence for a view in the strong sense of empirical evidence. Especially since our context is one of abstract metaphysics, when offering a thought-experiment as a refutation of some principle we should make sure that our case serves better as a reductio of that principle, rather than of one of our own assumptions. We have seen important examples drawn from the literature of all of these, as well as other, mistakes.

SECTION 2: *Looking further*

While many issues regarding personal identity and thought-experiments have been covered in the foregoing chapters, there remain some important points which have gone unmentioned, or which have received only scant attention. I wish to look briefly at some of those now, even if only to explain why I have not dealt with them or to indicate where I think further research might prove fruitful.

Perhaps the most notable gap is the lack of discussion regarding whose responses to thought-experiments are the ones which matter. The question which introduces a thought-experiment, as we have seen, is most commonly something of the form, "What would we say if..." But who is the "we" here? Much of my discussion has concerned <u>our</u> reactions to counterfactual situations, and the relative weighting of principles in <u>our</u> conceptual scheme. Again, to whom does "our" make reference? Is it everybody, people of a certain culture, academic philosophers, or who? The point is of some importance since different responses may well be forthcoming from different groups of "we's". If one wants a complete account of the philosophical knowledge which is to be gained from thought-experiments, this problem of whose responses count must be solved. I touched on this point in Chapter 6, but did little more than indicate one aspect of the problem.

I have left the question as to who "we" are an open one in the thesis. Doing this is in line with tradition in the debate on personal identity, but there are also further reasons for doing so. Because of the interconnectedness and open-endedness of philosophical problems, one has to draw some line at where to stop, or one can easily lead oneself from one problem into another, without end. To attempt an answer to the problem of who "we" are would either have required another thesis or would have been obviously inadequate.

A whole thesis would be required for the purpose, because the ramifications of the problem are so large. It would need a detailed discussion of the universality of concepts, the

possibility of conceptual objectivity and the relationship between circumstances and conceptual system. Closely related to these points is the hotly debated topic of the nature of rationality, which would also need discussion in order to achieve a satisfactory response to our problem. On the other hand, any quick attempt at a solution would come across as inadequate. Consider, for example, Unger's attempt to deal with the problem briefly:

> Finally a few words about whom to trust as respondents: To begin, each of us should, of course, take stock of our own intuitions, both on particular cases and on proposed generalities. In relation to a given subject area, however, some of us have intellectual investments that can strongly influence our responses. Indeed, just by having read a lot on the topic, one may become attached to a certain approach. Instead of responding in a way revelatory of one's untutored attitudes, one may then respond so as to favor an approach that, even if perhaps only temporarily, is conspicuously endorsed. Respondents who are so inclined are not to be much trusted. At the same time, a most useful respondent will rather fully understand even the more intricate offerings she encounters. Putting these two considerations together, our preferred respondents are those colleagues, and those students, who are as detached as they are astute. (Unger 1990: 13)

This leaves too many questions unanswered. Do we trust untutored intuitions, or those of students and teachers of the subject? Unger suggests both, although he also shows awareness of a tension here. Again, Unger's students at New

York University are worlds apart from, for example, my students at the University of Natal. Their backgrounds and the issues which concern them are very different. Should their responses to given situations clash, who do we take seriously? A quick-fix solution like Unger's just makes the problem more apparent. As a consequence, leaving the question open and as a subject for future research is a preferable alternative.

A second important problem which I have not faced squarely is the question of which thought-experiments we should take seriously and which we should ignore. Or rather, the problem not discussed is: what are permissable counterfactual scenarios? I have suggested no general way of distinguishing the permissable from the impermissible, although I have argued that specific situations are not to be taken seriously, and defended others against analogous attacks in the literature.

The aim of my thesis was to defend the method of thought-experiment, and to investigate via central examples in the literature what sort of things thought-experiments can establish. This does not strictly require an answer to the problem of which counterfactual scenarios are permissable, but there are reasons besides this for not attempting such an answer here. Once again, the central reason is that another thesis would be needed to do justice to the problem. We saw in Part One that attempts like Flew's and Wilkes's to provide the required distinguishing principles are inadequate. They suggest that only actual examples or cases which fit strictly

with the details of our scientific picture of the world are acceptable, but we saw good reasons for not drawing the line at such points. Perhaps there is indeed no clear line to be drawn between permissable and impermissible counterfactual situations.

The beginnings of a more sensible approach to the issue are to be found in Harre's outline of counterfactual epistemology (Harre 1983: 16-21)[1]. Harre suggests (as an extension to Lewis's [1973] theory on the semantics of counterfactuals) that our knowledge of the truth of a counterfactual conditional depends upon a set of categorical assertions related to the counterfactual. The more securely a counterfactual is grounded in reliable categorical assertions, the more secure we can be in its truth. But these are no more than the beginnings of the required account of counterfactual epistemology, and a fuller account must await further research.

One final area which is worth mention is clearly beyond the scope of this thesis, but is an area where the work here might well have useful applications. I have concerned myself with thought-experiments in the context of the personal identity debate, but they do not occur in that debate alone, nor are they (as is sometimes suggested) the preserve of philosophers. Work has been done on thought-experiments in science and elsewhere (Brown 1991, Sorensen 1992), but it would be extremely interesting to see how the points on thought-experiment methodology which have emerged here relate

to thought-experiments as they occur, for instance, in moral philosophy.

My discussion may be even more relevant to certain areas of law. For not only is the law greatly concerned with persons and matters which (like responsibility) depend on their identity, but thinly-disguised thought-experiments are frequently to be found in its domain. The law of delict is perhaps the most obviously related area. In delict, courts consider (amongst other things) what a reasonable person would have done in the circumstances, whether something would still have happened if X had not acted as he did, and what a person's life would have been like had they not been injured. In all of these cases, thought-experiments concerning persons and their identity are involved. It should also be noted that these are areas of the law in which courts often find it extremely difficult to reach decisions (Strassfield 1992). As a result, it would be interesting and important to see how philosophy might be of help here to the law[2].

For all these reasons, I believe that I have identified some things that thought-experiments can do, as well as some things which they cannot hope to achieve. At the same time I have pointed towards the large and complex terrain that still needs to be surveyed and mapped.

NOTES

1. A similar account is to be found in Rescher (1964).

2. I have begun a tentative investigation into these issues in "Counterfactuals and the Law" (Beck 1993b).

REFERENCES

Baillie,J (1990) "Identity, Survival and Sortal Concepts". Philosophical Quarterly 40, 183-194

Beck,S (1989) "Parfit and the Russians". Analysis 49, 205-209

(1991) "Finding Ourselves: Personal Identity and the Limits of Possible-World Arguments". South African Journal of Philosophy 10, 1-6

(1992a) "Should We Tolerate People Who Split?" Southern Journal of Philosophy 30, 1-17

(1992b) "The Method of Possible Worlds". Metaphilosophy 23, 119-131

(1993a) "Humans, Persons and Thought-Experiments". Journal for the Study of Religion 6, 31-53

(1993b) "Counterfactuals and the Law". South African Journal of Philosophy 12, 62-65

Block,N,ed. (1980) Readings in Philosophy of Psychology. London: Methuen

Butler,J (1736) "Of Personal Identity" in Perry (1975), 99-105

Care,NS & Grimm,RH,eds (1969) Perception and Personal Identity. Cleveland: Case Western Reserve University Press

Chisholm,R (1969) "The Loose and Popular and the Strict and Philosophical Senses of Identity" in Care and Grimm (1969), 82-106

Churchland,P (1981) "Eliminative Materialism and Propositional Attitudes". Journal of Philosophy 78, 67-90

Curley,EM (1982) "Leibniz on Locke on Personal Identity" in Hooker (1982), 303-326

Davidson,D (1970) "Mental Events" in Davidson (1980), 207-227

(1980) Essays on Actions and Events. Oxford: Oxford University Press

Dennett,D (1976) "Conditions of Personhood" in Dennett (1979), 267-285

(1979) Brainstorms. Hassocks: Harvester Press

(1981) "True Believers: The Intentional Strategy and Why It Works" in Heath (1981), 53-75

Devitt,M & Sterelny,K (1987) Language and Reality. Oxford: Blackwell

Elliot,R (1991) "Personal Identity and the Causal Continuity Requirement". Philosophical Quarterly 41, 55-75

Flew,A (1951) "Locke and the Problem of Personal Identity". Philosophy 26, 53-68

(1986) "Equality, yes surely, but Justice?" Philosophical Papers XV, 197-204

(1988) The Logic of Mortality. Oxford: Blackwell

Fodor,J (1964) "On Knowing What We Would Say". Philosophical Review 73, 198-212

Forbes,G (1985) The Metaphysics of Modality. Oxford: Clarendon Press

French,P (1983) "Kinds and Persons". Philosophy and Phenomenological Research 44, 241-254

Garrett,B (1986) "Possible Worlds and Personal Identity". Philosophical Books 27, 65-72

(1991) "Personal Identity and Reductionism". Philosophy and Phenomenological Research 51, 361-373

Gillett,G (1986) "Disembodied Persons". Philosophy 61, 377-386

Harre,R (1983) An Introduction to the Logic of the Sciences. London: Macmillan

Heath,AF,ed. (1981) Scientific Explanation. Oxford: Oxford University Press

Hooker,M,ed. (1982) Leibniz: Critical and Interpretive Essays. Minneapolis: University of Minnesota Press

Horgan,T & Woodward,J (1985) "Folk Psychology is Here To Stay". Philosophical Review 94, 197-226

Johnston,M (1987) "Human Beings". Journal of Philosophy 84, 59-83

Jolley,N (1984) Leibniz and Locke. Oxford: Clarendon Press

Kahneman,D & Tversky,A (1984) "Choices, Values and Frames". American Psychologist 39, 341-350

Kitcher,P (1979) "Natural Kinds and Unnatural Persons" Philosophy 54, 541-547

Kripke,S (1980) Naming and Necessity. Oxford: Blackwell

Leibniz,G (1765) New Essays on Human Understanding. Translated and edited by P Remnant and J Bennett. Cambridge: Cambridge University Press, 1981

Lewis,D (1972) "Psychophysical and Theoretical Identifications" in Block (1980), 207-215

(1973) Counterfactuals. Oxford: Blackwell

(1980) "Mad Pain and Martian Pain" in Block (1980), 216-222

Loar,B (1981) Mind and Meaning. Cambridge University Press

Locke,J (1694) "Of Identity and Diversity" in Perry (1975), 33-52

Lowe,EJ (1990) Review of H Noonan Personal Identity. Mind 99, 477-479

Mackie,P (1987) "Essence, Origin and Bare Identity". Mind 96, 173-201

Madell,G (1981) The Identity of the Self. Edinburgh": University Press

(1985) "Derek Parfit and Greta Garbo". Analysis 45, 105-109

Mates,B (1986) The Philosophy of Leibniz. New York: Oxford University Press

McTaggart,J (1927) The Nature of Existence Vol 2. Cambridge: Cambridge University Press

Noonan,H (1989) Personal Identity. London: Routledge and Kegan Paul

Nozick,R (1981) Philosophical Explanations. Oxford: Clarendon Press

Parfit,D (1984) Reasons and Persons. Oxford: Clarendon Press

(1976) "Lewis, Perry and What Matters" in Rorty (1976), 91-107

Perry,J,ed. (1975) Personal Identity. Berkeley: University of California Press

Putnam,H (1963) "It Ain't Necessarily So" in Rosenberg and Travis (1970), 52-62

(1964) "Robots: machines or artificially created life?" in Putnam (1981), 386-407

(1975) "The Meaning of 'Meaning'" in Putnam (1981), 215-271

(1980) "The Nature of Mental States" in Block (1980), 223-231

(1981) Mind, Language and Reality. Cambridge: Cambridge University Press

Quine,WvO (1972) Review of M.K.Munitz (ed) Identity and Individuation. Journal of Philosophy 69, 488-497

Rescher,N (1964) Hypothetical Reasoning. Amsterdam: North-Holland Publishing Co

(1991a) "Thought-Experimentation in Pre-Socratic Philosophy" in Rescher (1991b), 143-155

(1991b) Baffling Phenomena. Savage: Rowman and Littlefield

Rorty,A,ed. (1976) The Identities of Persons. Berkeley: University of California Press

Rosenberg,J & Travis,B,eds. (1970) Readings in the Philosophy of Language. Englewood Cliffs: Prentice-Hall

Russell,B (1905) "On Denoting" in Russell (1956), 39-56

(1956) Logic and Knowledge. London: Allen and Unwin

Sainsbury,RM (1989) Paradoxes. Cambridge: Cambridge University Press

Salmon,N (1982) Reference and Essence. Oxford: Blackwell

Searle,J (1958) "Proper Names". Mind 67, 166-173

Seddon,G (1972) "Logical Possibility". Mind 81, 481-494

Shoemaker,S (1963) Self-Knowledge and Self-Identity. New York: Cornell University Press

(1970) "Persons and Their Pasts" in Shoemaker (1984), 19-48

	(1984)	Identity, Cause and Mind. Cambridge: Cambridge University Press
Shoemaker,S & Swinburne,R	(1984)	Personal Identity. Oxford: Blackwell
Smart,JJC	(1959)	"Sensations and Brain Processes". Philosophical Review 68, 141-156
Sorensen,R	(1992)	Thought Experiments. Oxford: Oxford University Press
Stich,S	(1983)	From Folk Psychology to Cognitive Science. Cambridge, Mass: MIT Press
Strassfield,R	(1992)	"If...". George Washington Law Review 60, 339-416
Unger,P	(1982)	"Towards a Psychology of Common Sense". American Philosophical Quarterly 19, 117-129
	(1990)	Identity, Consciousness and Value. Oxford: Oxford University Press
White,S	(1989)	"Metapsychological Relativism and the Self" Journal of Philosophy 86, 298-323
Wiggins,D	(1967)	Identity and Spatio-Temporal Continuity. Oxford: Blackwell
	(1976)	"Locke, Butler and the Stream of Consciousness: and Men as a Natural Kind" in Rorty (1976), 139-173
	(1980)	Sameness and Substance. Oxford: Blackwell
Wilkes,K	(1988)	Real People. Oxford: Clarendon Press
Williams,B	(1966)	"Imagination and the Self" in Williams (1973), 26-45
	(1970)	"The Self and the Future" in Williams (1973), 46-63
	(1973)	Problems of the Self. Cambridge: Cambridge University Press
Wittgenstein	(1967)	Zettel (ed) G.E.M.Anscombe and G.H.von Wright. Oxford: Blackwell